```
*****************************************
* P6110: Statistical Computing with SAS *
*                                       *
* Homework 5                            *
* 10/27/2023                            *
* Name: Camille Okonkwo                 *
* UNI:  co2554                          *
*****************************************;

%let name = Camille Okonkwo;
%let date = 10/27/2023;
%let sharedrive = /home/u63212611/my_shared_file_links/u63093975;
%let localdrive = /home/u63212611/Statistical Computing with SAS;
%include "&sharedrive/P6110 Report Template.sas";

* ODS RTF OUTPUT;
options nodate nonumber leftmargin=1in rightmargin=1in colorprinting=yes;
* == BREAD == ; ods rtf
                file = "&localdrive/co2554_hw5_results.rtf"
                author="&name"
                startpage=never nogtitle nogfootnote image_dpi=300
                style=P6110;
* ==       == ; ods noproctitle; * supress procedure title;

** START OF DOCUMENT;

%DocTitle(Homework 5 Output);

%Qtitle(Q1. Child Care . . . Data Cleaning);

%Qsection(a. Import the dataset.);

proc format;
    value sitefmt   1 = "Urban"
                    2 = "Rural";
run;

*Import the dataset;
proc import out = childcare /*naming this dataset [childcare]*/
    datafile = "&sharedrive/Data/childcare.csv"
    dbms = csv replace;
    label   hhid    = "Housing ID"
            hours   = "Hours of child care (day)"
            days    = "Days of child care (day)"
            cost    = "Cost of child care (mth)"
            subsidy = "Child care subsidy (mth)"; *applying labels L5.1 to all subsequent questions;
run;

data childcare;
    set childcare;  *applying format F5.1 to site for all subsequent questions;
    format site sitefmt.;
run;

*Print the first 10 observations of [childcare].;
proc print data = childcare (obs = 10) label;
run;

%Qsection(b. Create a variable 'Survey');

*Create a variable Survey which shows the survey number for each participating household.;
data childcare;
    set childcare;
    by hhid; * Identifier ;
    retain Survey;
    if first.hhid then Survey = 1; * Initialize to 1 for the first survey in each household ;
    else Survey + 1; *Increment for subsequent surveys in the same household;
run;

*Create a table using PROC FREQ which shows the frequency of Survey by Site (no row or column percentages);
proc freq data = childcare;
    tables Survey*Site / nocol norow nopercent;
    format site sitefmt.;
run;

%newpage;
%Qnospace(Q2. Child Care . . . Baseline Data);

* Create a subset of [childcare] named [childcareBL], which contains all baseline survey responses.;
data childcareBL;
  set childcare;
  where wave = 1989;
run;

*Create a table using PROC TABULATE with [childcareBL];
proc tabulate   data = childcareBL;
    class site;
    var days hours cost subsidy;
    table
    (n median qrange)*(days hours cost subsidy),
    (site = '' ALL)
    / box = "1989 Housing Surveys";
    keylabel ALL = "All Households" qrange = "IQR";
run;

/*
Describe your findings:
At the baseline year of the China Health and Nutrition Study
(1989), there were more rural than urban households participating
in the survey on household child care discussing daily hours,
days per week, and monthly cost of non-household child care,
and monthly government subsidy for child care costs. The IQR
for rural household's days of childcare (6 days) is wider than urban
households (5 days), suggesting more variabilty in the number of days
for care. Urban households have a higher median number of hours (8 hours), but
their IQR is narrower (3 hours), indicating less variability. Urban households
have a greater range of costs, with their IQR 47.5 yuan of and median of 20 yuan.
Urban households receive a higher median subsidy, and their IQR is also wider,
indicating varying subsidy levels.
*/


ods text= "%bold(Describe your findings:) At the baseline year of the China Health and Nutrition Study
(1989), there were more rural than urban households participating in the survey on household child care discussing
days per week, and monthly cost of non-household child care, and monthly government subsidy for child care costs. T
for rural household's days of childcare (6 days) is wider than urban households (5 days), suggesting more variabilt
for care. Urban households have a higher median number of hours (8 hours), but their IQR is narrower (3 hours), ind
have a greater range of costs, with their IQR 47.5 yuan of and median of 20 yuan. Urban households receive a higher
indicating a variation in subsidy levels.";
%newpage;
%Qnospace(Q3. Refugee Appeals . . . Messy Longitudinal Plots);

%Qsection(a. Create a series plot of cost of child care over time by site using [childcare]);
*Subsetting data to exclude missing cost rows;
/* Step 1: Data Preparation - Exclude missing data for cost and households with only one survey response */

data childcare_filtered;
    set childcare;
    where not missing(cost);
    format site sitefmt.;
run;

data childcare_filtered;
    set childcare_filtered;
    by hhid;
    if Survey > 1 then do;
    flag = 1;
    end;
    if last.hhid and flag = "." then do;
    delete;
    end;
run;

data childcare_plot;
    set childcare_filtered;
    by hhid; output;
    if last.hhid then do;
    cost = .; output;
end;
run;

*Creating series plot;
proc sgplot data=childcare_plot;
    series x=wave y=cost
    /markers group=site break;
    xaxis label="Year";
    yaxis label="Cost of child care (Yuan)";
    format site sitefmt.;
run;


/*
```

```
This an awful plot. There are too many lines to clearly and association between cost and wave? What features are
good? What features are bad?
*/

ods text= "%bold(Describe your whether you think this is a good figure:) "

%Qsection(b. Create a plot of regression lines for cost of child care over time by site using
[childcare] with the same data exclusions as Q3.A.);
proc sgplot data=childcare_filtered;
  reg y=cost x=Wave / group=Site lineattrs=(thickness=1) legendlabel="Site";
  xaxis label="Year";
  yaxis label="Cost of Child Care (yuan)";
  keylegend / location=inside position=topright;
  title "Regression Lines for Cost of Child Care Over Time by Site";
run;
/*
This is a better because the data points are less clustered together.
*/

ods text= "%bold(Describe your whether you think this is a good figure:) This is a better because the data points a

%newpage;
%Qnospace(Q4. Child Care . . . Nice Longitudinal Plot);

*Create a plot for the average monthly child care cost per month over time by site.;
proc means      data = childcare noprint;
    class site wave;
    var cost;
    output out = avg_childcare mean = / autoname; *calculating average month childcare costs;
run;
data missing_childcare;
    set avg_childcare;
    if missing (site) or missing (wave) then delete;
    samplesize = cat ("n=", _freq_);
run;
proc sgplot     data = missing_childcare;
    pbspline x = wave y = cost_mean /*connecting by spline*/
    / group = site datalabel = samplesize;
    xaxis label = "Year";
    yaxis label = "Cost of child care (Yuan)";
run;


/*
Describe your findings:
This plot is better than the one created in Q3 because of the lack of noise in the data.
You can clearly see the trendlines, and N over time.
*/

ods text= "%bold(Describe your findings:) This plot is better than the one created in Q3 because of the lack of noi
You can clearly see the trendlines, and N over time.";

%newpage;
%Qnospace(Q5. (Bonus). More Longitudinal Plots!);



* == BREAD == ; ods rtf close;
```