```
1          OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
68
69         ************************************************************************
70         * Process Data Script
71         * Murphy John
72         * 2025-04-07
73         * This script loads, processes, and compiles the data used in this project.
74         ************************************************************************;
75
76         title "Setup";
77
78         ** footnote;
79         footnote "Data processing script run on &SYSDATE at &SYSTIME.";
80
81         ** establish library;
82         libname mylib "/home/u63984496/BIOS7400/final-project";
NOTE: Libref MYLIB was successfully assigned as follows:
      Engine:        V9
      Physical Name: /home/u63984496/BIOS7400/final-project
83         ************************************************************************;
84
85         title "Data processing";
86
87         title2 "Avocado Data";
88
89         * load data;
90         proc import datafile="/home/u63984496/BIOS7400/final-project/avocado.csv"
91         out=work.raw_avo
92         dbms=csv
93         replace;
94         guessingrows=MAX;
95         run;

NOTE: Unable to open parameter catalog: SASUSER.PARMS.PARMS.SLIST in update mode. Temporary parameter values will be saved to
WORK.PARMS.PARMS.SLIST.
Name  is not a valid SAS name.
Problems were detected with provided names.  See LOG.
96          /*---------------------------------------------------------------------
97          *   PRODUCT:   SAS
98          *   VERSION:   9.4
99          *   CREATOR:   External File Interface
100         *   DATE:      30APR25
101         *   DESC:      Generated SAS Datastep Code
102         *   TEMPLATE SOURCE:  (None Specified.)
103         ---------------------------------------------------------------------*/
104            data WORK.RAW_AVO    ;
105            %let _EFIERR_ = 0; /* set the ERROR detection macro variable */
106            infile '/home/u63984496/BIOS7400/final-project/avocado.csv' delimiter = ',' MISSOVER DSD lrecl=32767 firstobs=2 ;
107               informat VAR1 best32. ;
108               informat Date yymmdd10. ;
109               informat AveragePrice best32. ;
110               informat "Total Volume"N best32. ;
111               informat "4046"N best32. ;
112               informat "4225"N best32. ;
113               informat "4770"N best32. ;
114               informat "Total Bags"N best32. ;
115               informat "Small Bags"N best32. ;
116               informat "Large Bags"N best32. ;
117               informat "XLarge Bags"N best32. ;
118               informat type $12. ;
119               informat year best32. ;
120               informat region $19. ;
121               format VAR1 best12. ;
122               format Date yymmdd10. ;
123               format AveragePrice best12. ;
124               format "Total Volume"N best12. ;
125               format "4046"N best12. ;
126               format "4225"N best12. ;
127               format "4770"N best12. ;
128               format "Total Bags"N best12. ;
129               format "Small Bags"N best12. ;
130               format "Large Bags"N best12. ;
131               format "XLarge Bags"N best12. ;
132               format type $12. ;
133               format year best12. ;
134               format region $19. ;
135            input
136                        VAR1
137                        Date
138                        AveragePrice
139                        "Total Volume"N
140                        "4046"N
```

```
141                        "4225"N
142                        "4770"N
143                        "Total Bags"N
144                        "Small Bags"N
145                        "Large Bags"N
146                        "XLarge Bags"N
147                        type  $
148                        year
149                        region  $
150            ;
151            if _ERROR_ then call symputx('_EFIERR_',1);  /* set ERROR detection macro variable */
152            run;

NOTE: The infile '/home/u63984496/BIOS7400/final-project/avocado.csv' is:
      Filename=/home/u63984496/BIOS7400/final-project/avocado.csv,
      Owner Name=u63984496,Group Name=oda,
      Access Permission=-rw-r--r--,
      Last Modified=21Apr2025:11:51:20,
      File Size (bytes)=1989197

NOTE: 18249 records were read from the infile '/home/u63984496/BIOS7400/final-project/avocado.csv'.
      The minimum record length was 77.
      The maximum record length was 135.
NOTE: The data set WORK.RAW_AVO has 18249 observations and 14 variables.
NOTE: DATA statement used (Total process time):
      real time           0.02 seconds
      user cpu time       0.02 seconds
      system cpu time     0.00 seconds
      memory              12047.78k
      OS Memory           36384.00k
      Timestamp           04/30/2025 05:39:50 PM
      Step Count                        299  Switch Count  2
      Page Faults                       0
      Page Reclaims                     426
      Page Swaps                        0
      Voluntary Context Switches        17
      Involuntary Context Switches      1
      Block Input Operations            0
      Block Output Operations           4624


18249 rows created in WORK.RAW_AVO from /home/u63984496/BIOS7400/final-project/avocado.csv.



NOTE: WORK.RAW_AVO data set was successfully created.
NOTE: The data set WORK.RAW_AVO has 18249 observations and 14 variables.
NOTE: PROCEDURE IMPORT used (Total process time):
      real time           6.82 seconds
      user cpu time       6.71 seconds
      system cpu time     0.03 seconds
      memory              12047.78k
      OS Memory           36640.00k
      Timestamp           04/30/2025 05:39:50 PM
      Step Count                        299  Switch Count  9
      Page Faults                       0
      Page Reclaims                     7523
      Page Swaps                        0
      Voluntary Context Switches        139
      Involuntary Context Switches      46
      Block Input Operations            0
      Block Output Operations           4720


153
154        * data processing;
155        data work.clean_avo;
156        * read raw avocado data;
157        * rename select variables;
158        set work.raw_avo(rename = (
159        AveragePrice = avgprice
160        'Total Volume'n = totvol
161        '4046'n = totsm
162        '4225'n = totlg
163        '4770'n = totxl
164        'Total Bags'n = totbags
165        'Small Bags'n = totbags_sm
166        'Large Bags'n = totbags_lg
167        'XLarge Bags'n = totbags_xl
168        ));
169
170        * seperate date by month and year;
171        * create a month year variable;
172        month = put(date, monname.);
```

```
173         month_num = month(date);
174         month = strip(propcase(month));
175         date = mdy(month_num, 1, year);
176
177         * keep only specifc regions;
178         if region not in (
179                 "California",
180                 "West",
181                 "Northeast",
182                 "SouthCentral",
183                 "Southeast",
184                 "GreatLakes",
185                 "MidSouth",
186                 "Plains")
187                 then delete;
188
189         drop VAR1;
190         run;

NOTE: There were 18249 observations read from the data set WORK.RAW_AVO.
NOTE: The data set WORK.CLEAN_AVO has 2366 observations and 15 variables.
NOTE: DATA statement used (Total process time):
      real time           0.01 seconds
      user cpu time       0.01 seconds
      system cpu time     0.00 seconds
      memory              2746.68k
      OS Memory           30636.00k
      Timestamp           04/30/2025 05:39:50 PM
      Step Count                        300  Switch Count  2
      Page Faults                       0
      Page Reclaims                     287
      Page Swaps                        0
      Voluntary Context Switches        12
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           776


191
192         * group by year, month, region, and type;
193         proc sql;
194             create table work.avo_group as
195             select
196                 year,
197                 month,
198                 month_num,
199                 date,
200                 region,
201                 type,
202                 mean(avgprice) as avgprice format=8.2,
203                 sum(totvol) as totvol,
204                 sum(totsm) as totsm,
205                 sum(totlg) as totlg,
206                 sum(totxl) as totxl,
207                 sum(totbags) as totbags,
208                 sum(totbags_sm) as totbags_sm,
209                 sum(totbags_lg) as totbags_lg,
210                 sum(totbags_xl) as totbags_xl
211             from work.clean_avo
212             group by date, region, type;
NOTE: The query requires remerging summary statistics back with the original data.
NOTE: Table WORK.AVO_GROUP created, with 2366 rows and 15 columns.

213         quit;
NOTE: PROCEDURE SQL used (Total process time):
      real time           0.00 seconds
      user cpu time       0.01 seconds
      system cpu time     0.01 seconds
      memory              7315.60k
      OS Memory           34860.00k
      Timestamp           04/30/2025 05:39:50 PM
      Step Count                        301  Switch Count  2
      Page Faults                       0
      Page Reclaims                     370
      Page Swaps                        0
      Voluntary Context Switches        27
      Involuntary Context Switches      1
      Block Input Operations            0
      Block Output Operations           784


214
215         * sort by date and remove duplicate obs;
216         proc sort data=work.avo_group nodupkey out=work.dat_avo;
```

```
217           by date region type;
218       run;

NOTE: There were 2366 observations read from the data set WORK.AVO_GROUP.
NOTE: 1820 observations with duplicate key values were deleted.
NOTE: The data set WORK.DAT_AVO has 546 observations and 15 variables.
NOTE: PROCEDURE SORT used (Total process time):
      real time           0.00 seconds
      user cpu time       0.00 seconds
      system cpu time     0.00 seconds
      memory              2725.31k
      OS Memory           31292.00k
      Timestamp           04/30/2025 05:39:50 PM
      Step Count                        302  Switch Count  2
      Page Faults                       0
      Page Reclaims                     240
      Page Swaps                        0
      Voluntary Context Switches        12
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           272


219
220       * print first 10 obs;
221       proc print data=work.dat_avo(obs=10);
222       run;

NOTE: There were 10 observations read from the data set WORK.DAT_AVO.
NOTE: PROCEDURE PRINT used (Total process time):
      real time           0.02 seconds
      user cpu time       0.03 seconds
      system cpu time     0.00 seconds
      memory              1619.96k
      OS Memory           29608.00k
      Timestamp           04/30/2025 05:39:50 PM
      Step Count                        303  Switch Count  0
      Page Faults                       0
      Page Reclaims                     62
      Page Swaps                        0
      Voluntary Context Switches        0
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           0


223
224       title2 "Temperature Data";
225
226       ** load data;
227       filename raw_temp '/home/u63984496/BIOS7400/final-project/temp.txt';
228       data dat_temp;
229       * read raw temp data;
230       infile raw_temp;
231
232       * use absolute input pointer control;
233       input @;
234
235       * delete non-numeric values;
236       if notdigit(scan(_infile_, 1)) then delete;
237
238       * create year and month columns;
239       else input year January February March April May June July August September October November December;
240
241       * keep only years 2015 - 2018;
242       if year < 2015 or year > 2018 then delete;
243
244       * temperatures are in 0.01 degrees C. convert to actual degrees C;
245       * pivot longer to create a month/year column and temp column;
246       length month $9;
247       array col{12} January February March April May June July August September October November December;
248       do i = 1 to 12;
249       temp = round(col{i} / 100, 0.01);
250       month = vname(col{i});
251       output;
252       end;
253       month = strip(propcase(month));
254
255       * keep year month temp cols only;
256       keep year month temp
257
258       run;
259
260       * print first 10 obs;
```

```
WARNING: The variable run in the DROP, KEEP, or RENAME list has never been referenced.
NOTE: The infile RAW_TEMP is:
      Filename=/home/u63984496/BIOS7400/final-project/temp.txt,
      Owner Name=u63984496,Group Name=oda,
      Access Permission=-rw-r--r--,
      Last Modified=29Apr2025:13:23:13,
      File Size (bytes)=16938

NOTE: Invalid data for April in line 168 22-25.
NOTE: Invalid data for May in line 168 27-30.
NOTE: Invalid data for June in line 168 32-35.
NOTE: Invalid data for July in line 168 37-40.
NOTE: Invalid data for August in line 168 42-45.
NOTE: Invalid data for September in line 168 47-50.
NOTE: Invalid data for October in line 168 52-55.
NOTE: Invalid data for November in line 168 57-60.
NOTE: Invalid data for December in line 168 62-65.
RULE:     ----+----1----+----2----+----3----+----4----+----5----+----6----+----7----+----8----+----9----+----0
168       2025   183  162  182 **** **** **** **** **** **** **** **** ****   **** ***    172 **** **** ****
     101  2025 104
year=2025 January=183 February=162 March=182 April=. May=. June=. July=. August=. September=. October=. November=. December=.
month=  i=. temp=. _ERROR_=1
_INFILE_=2025   183  162  182 **** **** **** **** **** **** **** **** ****   **** ***    172 **** **** ****  2025 _N_=168
NOTE: 175 records were read from the infile RAW_TEMP.
      The minimum record length was 0.
      The maximum record length was 104.
NOTE: The data set WORK.DAT_TEMP has 48 observations and 3 variables.
NOTE: DATA statement used (Total process time):
      real time           0.00 seconds
      user cpu time       0.01 seconds
      system cpu time     0.00 seconds
      memory              1020.50k
      OS Memory           29608.00k
      Timestamp           04/30/2025 05:39:50 PM
      Step Count                        304  Switch Count  2
      Page Faults                       0
      Page Reclaims                     90
      Page Swaps                        0
      Voluntary Context Switches        18
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           264


261       proc print data=work.dat_temp(obs=10);
262       run;

NOTE: There were 10 observations read from the data set WORK.DAT_TEMP.
NOTE: PROCEDURE PRINT used (Total process time):
      real time           0.01 seconds
      user cpu time       0.01 seconds
      system cpu time     0.00 seconds
      memory              774.12k
      OS Memory           29608.00k
      Timestamp           04/30/2025 05:39:50 PM
      Step Count                        305  Switch Count  0
      Page Faults                       0
      Page Reclaims                     62
      Page Swaps                        0
      Voluntary Context Switches        0
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           24


263
264       title2 "President Data";
265
266       *** In 2015 and 2016, Barack Obama of the democratic party was president of the US.
267       *** In 2017 and 2018, Donald Trump of the republican party was president of the US.;
268
269       * establish data;
270       data dat_pres;
271       length year 4 president $ 20 pres_party $ 25;
272           input year president pres_party;
273           infile datalines dsd dlm = " ";
274           datalines;

NOTE: The data set WORK.DAT_PRES has 4 observations and 3 variables.
NOTE: DATA statement used (Total process time):
      real time           0.00 seconds
      user cpu time       0.00 seconds
      system cpu time     0.00 seconds
```

```
      memory              668.34k
      OS Memory           29608.00k
      Timestamp           04/30/2025 05:39:50 PM
      Step Count                        306  Switch Count  2
      Page Faults                       0
      Page Reclaims                     85
      Page Swaps                        0
      Voluntary Context Switches        11
      Involuntary Context Switches      1
      Block Input Operations            0
      Block Output Operations           264


279         ;
280         run;
281
282         * print;
283         proc print data=work.dat_pres;
284         run;

NOTE: There were 4 observations read from the data set WORK.DAT_PRES.
NOTE: PROCEDURE PRINT used (Total process time):
      real time           0.00 seconds
      user cpu time       0.01 seconds
      system cpu time     0.00 seconds
      memory              714.53k
      OS Memory           29608.00k
      Timestamp           04/30/2025 05:39:50 PM
      Step Count                        307  Switch Count  0
      Page Faults                       0
      Page Reclaims                     62
      Page Swaps                        0
      Voluntary Context Switches        0
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           0


285         ********************************************************************************;
286
287         title "Data merging";
288         * sql can handle many-to-one merging;
289         * save to mylib;
290         proc sql;
291             create table work.dat_merge as
292             select
293                 a.*,
294                 b.*,
295                 c.*
296             from work.dat_avo as a
297             inner join work.dat_temp as b
298             on a.year = b.year and a.month = b.month
299             inner join work.dat_pres as c
300                 on a.year = c.year;
WARNING: Variable year already exists on file WORK.DAT_MERGE.
WARNING: Variable month already exists on file WORK.DAT_MERGE.
WARNING: Variable year already exists on file WORK.DAT_MERGE.
NOTE: Table WORK.DAT_MERGE created, with 546 rows and 18 columns.

301         quit;
NOTE: PROCEDURE SQL used (Total process time):
      real time           0.00 seconds
      user cpu time       0.00 seconds
      system cpu time     0.01 seconds
      memory              6343.62k
      OS Memory           35252.00k
      Timestamp           04/30/2025 05:39:50 PM
      Step Count                        308  Switch Count  7
      Page Faults                       0
      Page Reclaims                     174
      Page Swaps                        0
      Voluntary Context Switches        25
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           272


302
303         * add labels to variables;
304         data mylib.dat;
305         set work.dat_merge;
306         label
307         year = "Year"
308         month = "Month Name"
```

```
309        month_num = "Month Number"
310        date = "Date of observation- only month and years are known"
311        region = "City or region of the observation"
312        type = "Type of farming method"
313        avgprice = "Average price of a single avocado"
314        totvol = "Total Number of avocados sold"
315        totsm = "Total number of avocados with PLU 4046 (small) sold"
316        totlg = "Total number of avocados with PLU 4225 (large) sold"
317        totxl = "Total number of avocados with PLU 4770 (xlarge) sold"
318        totbags = "Total number of bags sold"
319        totbags_sm = "Total number of PLU 4046 (small) bags sold"
320        totbags_lg = "Total number of PLU 4225 (large) bags sold"
321        totbags_xl = "Total number of PLU 4770 (xlarge) bags sold"
322        temp = "Temperature difference (degress C)"
323        president = "Name of current U.S. president"
324        pres_party = "Poliical Party of current U.S. president";
325        run;

NOTE: There were 546 observations read from the data set WORK.DAT_MERGE.
NOTE: The data set MYLIB.DAT has 546 observations and 18 variables.
NOTE: DATA statement used (Total process time):
      real time           0.01 seconds
      user cpu time       0.00 seconds
      system cpu time     0.00 seconds
      memory              1188.65k
      OS Memory           29868.00k
      Timestamp           04/30/2025 05:39:50 PM
      Step Count                        309  Switch Count  1
      Page Faults                       0
      Page Reclaims                     94
      Page Swaps                        0
      Voluntary Context Switches        35
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           264


326        ********************************************************************************;
327
328        title "Print data";
329
330        * print first 10 obs;
331        proc print data=mylib.dat(obs=10);
332        run;

NOTE: There were 10 observations read from the data set MYLIB.DAT.
NOTE: PROCEDURE PRINT used (Total process time):
      real time           0.02 seconds
      user cpu time       0.03 seconds
      system cpu time     0.00 seconds
      memory              853.62k
      OS Memory           29608.00k
      Timestamp           04/30/2025 05:39:51 PM
      Step Count                        310  Switch Count  0
      Page Faults                       0
      Page Reclaims                     62
      Page Swaps                        0
      Voluntary Context Switches        9
      Involuntary Context Switches      1
      Block Input Operations            0
      Block Output Operations           16


333
334        * get frequency tables;
335        proc freq data=mylib.dat;
336        tables year month region type pres_party;
337        run;

NOTE: There were 546 observations read from the data set MYLIB.DAT.
NOTE: PROCEDURE FREQ used (Total process time):
      real time           0.04 seconds
      user cpu time       0.04 seconds
      system cpu time     0.00 seconds
      memory              1127.68k
      OS Memory           29868.00k
      Timestamp           04/30/2025 05:39:51 PM
      Step Count                        311  Switch Count  2
      Page Faults                       0
      Page Reclaims                     129
      Page Swaps                        0
      Voluntary Context Switches        22
      Involuntary Context Switches      2
      Block Input Operations            0
```

```
      Block Output Operations           280


338
339         * describe dataset;
340         proc contents data=mylib.dat;
341         run;

NOTE: PROCEDURE CONTENTS used (Total process time):
      real time           0.03 seconds
      user cpu time       0.04 seconds
      system cpu time     0.00 seconds
      memory              1069.00k
      OS Memory           29868.00k
      Timestamp           04/30/2025 05:39:51 PM
      Step Count                        312  Switch Count  0
      Page Faults                       0
      Page Reclaims                     102
      Page Swaps                        0
      Voluntary Context Switches        7
      Involuntary Context Switches      2
      Block Input Operations            0
      Block Output Operations           24


342
343         * END OF SCRIPT;
344
345         OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
355
```