```
1          OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
NOTE: ODS statements in the SAS Studio environment may disable some output features.
69         
70         * Aggregated Data Scripts
71         * Murphy John;
72         
73         
74         ****************************************************************************
75         * Process Data Script
76         * Murphy John
77         * 2025-04-07
78         * This script loads, processes, and compiles the data used in this project.
79         ****************************************************************************;
80         
81         title "Setup";
82         
83         ** footnote;
84         footnote "Data processing script run on &SYSDATE at &SYSTIME.";
85         
86         ** establish library;
87         libname mylib "/home/u63984496/BIOS7400/final-project";
NOTE: Libref MYLIB was successfully assigned as follows: 
      Engine:        V9 
      Physical Name: /home/u63984496/BIOS7400/final-project
88         ****************************************************************************;
89         
90         title "Data processing";
91         
92         title2 "Avocado Data";
93         
94         * load data;
95         proc import datafile="/home/u63984496/BIOS7400/final-project/avocado.csv"
96         out=work.raw_avo
97         dbms=csv
98         replace;
99         guessingrows=MAX;
100        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.
101         /**********************************************************************
102          *   PRODUCT:   SAS
103          *   VERSION:   9.4
104          *   CREATOR:   External File Interface
105          *   DATE:      01MAY25
106          *   DESC:      Generated SAS Datastep Code
107          *   TEMPLATE SOURCE:  (None Specified.)
108          ***********************************************************************/
109             data WORK.RAW_AVO    ;
110             %let _EFIERR_ = 0; /* set the ERROR detection macro variable */
111             infile '/home/u63984496/BIOS7400/final-project/avocado.csv' delimiter = ',' MISSOVER DSD lrecl=32767 firstobs=2 ;
112                informat VAR1 best32. ;
113                informat Date yymmdd10. ;
114                informat AveragePrice best32. ;
115                informat "Total Volume"N best32. ;
116                informat "4046"N best32. ;
117                informat "4225"N best32. ;
118                informat "4770"N best32. ;
119                informat "Total Bags"N best32. ;
120                informat "Small Bags"N best32. ;
121                informat "Large Bags"N best32. ;
122                informat "XLarge Bags"N best32. ;
123                informat type $12. ;
124                informat year best32. ;
125                informat region $19. ;
126                format VAR1 best12. ;
127                format Date yymmdd10. ;
128                format AveragePrice best12. ;
129                format "Total Volume"N best12. ;
130                format "4046"N best12. ;
131                format "4225"N best12. ;
132                format "4770"N best12. ;
133                format "Total Bags"N best12. ;
134                format "Small Bags"N best12. ;
135                format "Large Bags"N best12. ;
136                format "XLarge Bags"N best12. ;
137                format type $12. ;
138                format year best12. ;
139                format region $19. ;
140             input
```

```
                    VAR1
                    Date
                    AveragePrice
                    "Total Volume"N
                    "4046"N
                    "4225"N
                    "4770"N
                    "Total Bags"N
                    "Small Bags"N
                    "Large Bags"N
                    "XLarge Bags"N
                    type  $
                    year
                    region  $
        ;
        if _ERROR_ then call symputx('_EFIERR_',1);  /* set ERROR detection macro variable */
        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              12072.37k
      OS Memory           32544.00k
      Timestamp           05/01/2025 05:24:23 PM
      Step Count                        24  Switch Count  2
      Page Faults                       0
      Page Reclaims                     458
      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           6.79 seconds
      user cpu time       6.67 seconds
      system cpu time     0.03 seconds
      memory              12072.37k
      OS Memory           32800.00k
      Timestamp           05/01/2025 05:24:23 PM
      Step Count                        24  Switch Count  10
      Page Faults                       0
      Page Reclaims                     8060
      Page Swaps                        0
      Voluntary Context Switches        144
      Involuntary Context Switches      30
      Block Input Operations            0
      Block Output Operations           4720



    * data processing;
    data work.clean_avo;
    * read raw avocado data;
    * rename select variables;
    set work.raw_avo(rename = (
    AveragePrice = avgprice
    'Total Volume'n = totvol
    '4046'n = totsm
    '4225'n = totlg
    '4770'n = totxl
    'Total Bags'n = totbags
    'Small Bags'n = totbags_sm
    'Large Bags'n = totbags_lg
    'XLarge Bags'n = totbags_xl
173         ));
174
175         * seperate date by month and year;
176         * create a month year variable;
177         month = put(date, monname.);
178         month_num = month(date);
179         month = strip(propcase(month));
180         date = mdy(month_num, 1, year);
181
182         * keep only specififc regions;
183         if region not in (
184                 "California",
185                 "West",
186                 "Northeast",
187                 "SouthCentral",
188                 "Southeast",
189                 "GreatLakes",
190                 "MidSouth",
191                 "Plains")
192                 then delete;
193
194         drop VAR1;
195         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              2578.09k
      OS Memory           26540.00k
      Timestamp           05/01/2025 05:24:23 PM
      Step Count                        25  Switch Count  2
      Page Faults                       0
      Page Reclaims                     294
      Page Swaps                        0
      Voluntary Context Switches        10
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           776


196
197         * group by year, month, region, and type;
198         proc sql;
199             create table work.avo_group as
200             select
201                 year,
202                 month,
203                 month_num,
204                 date,
205                 region,
206                 type,
207                 mean(avgprice) as avgprice format=8.2,
208                 sum(totvol) as totvol,
209                 sum(totsm) as totsm,
210                 sum(totlg) as totlg,
211                 sum(totxl) as totxl,
212                 sum(totbags) as totbags,
213                 sum(totbags_sm) as totbags_sm,
214                 sum(totbags_lg) as totbags_lg,
215                 sum(totbags_xl) as totbags_xl
216             from work.clean_avo
217             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.

218         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              7264.57k
      OS Memory           31800.00k
      Timestamp           05/01/2025 05:24:23 PM
      Step Count                        26  Switch Count  2
      Page Faults                       0
      Page Reclaims                     472
      Page Swaps                        0
      Voluntary Context Switches        28
      Involuntary Context Switches      1
      Block Input Operations            0
      Block Output Operations           784
```

```

           * sort by date and remove duplicate obs;
           proc sort data=work.avo_group nodupkey out=work.dat_avo;
               by date region type;
           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              2611.93k
      OS Memory           28732.00k
      Timestamp           05/01/2025 05:24:23 PM
      Step Count                        27  Switch Count  2
      Page Faults                       0
      Page Reclaims                     259
      Page Swaps                        0
      Voluntary Context Switches        10
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           280



           * print first 10 obs;
           proc print data=work.dat_avo(obs=10);
           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              1608.81k
      OS Memory           27048.00k
      Timestamp           05/01/2025 05:24:23 PM
      Step Count                        28  Switch Count  0
      Page Faults                       0
      Page Reclaims                     284
      Page Swaps                        0
      Voluntary Context Switches        0
      Involuntary Context Switches      2
      Block Input Operations            0
      Block Output Operations           16



           title2 "Temperature Data";

           ** load data;
           filename raw_temp '/home/u63984496/BIOS7400/final-project/temp.txt';
           data dat_temp;
           * read raw temp data;
           infile raw_temp;

           * use absolute input pointer control;
           input @;

           * delete non-numeric values;
           if notdigit(scan(_infile_, 1)) then delete;

           * create year and month columns;
           else input year January February March April May June July August September October November December;

           * keep only years 2015 - 2018;
           if year < 2015 or year > 2018 then delete;

           * temperatures are in 0.01 degrees C. convert to actual degrees C;
           * pivot longer to create a month/year column and temp column;
           length month $9;
           array col{12} January February March April May June July August September October November December;
           do i = 1 to 12;
           temp = round(col{i} / 100, 0.01);
           month = vname(col{i});
           output;
           end;
           month = strip(propcase(month));

           * keep year month temp cols only;
```

```
261        keep year month temp
262
263        run;
264
265        * 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       0.00 seconds
      system cpu time     0.00 seconds
      memory              863.09k
      OS Memory           27048.00k
      Timestamp           05/01/2025 05:24:23 PM
      Step Count                        29  Switch Count  2
      Page Faults                       0
      Page Reclaims                     91
      Page Swaps                        0
      Voluntary Context Switches        18
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           272


266        proc print data=work.dat_temp(obs=10);
267        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       0.02 seconds
      system cpu time     0.00 seconds
      memory              662.43k
      OS Memory           27048.00k
      Timestamp           05/01/2025 05:24:23 PM
      Step Count                        30  Switch Count  0
      Page Faults                       0
      Page Reclaims                     77
      Page Swaps                        0
      Voluntary Context Switches        0
      Involuntary Context Switches      3
      Block Input Operations            0
      Block Output Operations           16


268
269        title2 "President Data";
270
271        *** In 2015 and 2016, Barack Obama of the democratic party was president of the US.
272        *** In 2017 and 2018, Donald Trump of the republican party was president of the US.;
273
274        * establish data;
275        data dat_pres;
276        length year 4 president $ 20 pres_party $ 25;
277            input year president pres_party;
278            infile datalines dsd dlm = " ";
279            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              668.28k
      OS Memory           27048.00k
      Timestamp           05/01/2025 05:24:23 PM
      Step Count                        31  Switch Count  2
      Page Faults                       0
      Page Reclaims                     85
      Page Swaps                        0
      Voluntary Context Switches        17
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           264
      

284        ;
285        run;
286
287        * print;
288        proc print data=work.dat_pres;
289        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              606.15k
      OS Memory           27048.00k
      Timestamp           05/01/2025 05:24:23 PM
      Step Count                        32  Switch Count  0
      Page Faults                       0
      Page Reclaims                     63
      Page Swaps                        0
      Voluntary Context Switches        0
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           0
      

290        ****************************************************************************;
291
292        title "Data merging";
293        * sql can handle many-to-one merging;
294        * save to mylib;
295        proc sql;
296            create table work.dat_merge as
297            select
298                a.*,
299                b.*,
300                c.*
301            from work.dat_avo as a
302            inner join work.dat_temp as b
303            on a.year = b.year and a.month = b.month
304            inner join work.dat_pres as c
305                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.

306        quit;
NOTE: PROCEDURE SQL used (Total process time):
      real time           0.00 seconds
      user cpu time       0.00 seconds
      system cpu time     0.00 seconds
      memory              6221.53k
      OS Memory           32692.00k
      Timestamp           05/01/2025 05:24:23 PM
      Step Count                        33  Switch Count  6
      Page Faults                       0
      Page Reclaims                     176
      Page Swaps                        0
      Voluntary Context Switches        23
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           272
      

307
308        * add labels to variables;
```

```
309        data mylib.dat;
310        set work.dat_merge;
311        label
312        year = "Year"
313        month = "Month Name"
314        month_num = "Month Number"
315        date = "Date of observation- only month and years are known"
316        region = "City or region of the observation"
317        type = "Type of farming method"
318        avgprice = "Average price of a single avocado"
319        totvol = "Total Number of avocados sold"
320        totsm = "Total number of avocados with PLU 4046 (small) sold"
321        totlg = "Total number of avocados with PLU 4225 (large) sold"
322        totxl = "Total number of avocados with PLU 4770 (xlarge) sold"
323        totbags = "Total number of bags sold"
324        totbags_sm = "Total number of PLU 4046 (small) bags sold"
325        totbags_lg = "Total number of PLU 4225 (large) bags sold"
326        totbags_xl = "Total number of PLU 4770 (xlarge) bags sold"
327        temp = "Temperature difference (degress C)"
328        president = "Name of current U.S. president"
329        pres_party = "Poliical Party of current U.S. president";
330        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              976.50k
      OS Memory           27308.00k
      Timestamp           05/01/2025 05:24:23 PM
      Step Count                        34  Switch Count  1
      Page Faults                       0
      Page Reclaims                     94
      Page Swaps                        0
      Voluntary Context Switches        39
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           264


331        ****************************************************************************;
332
333        title "Print data";
334
335        * print first 10 obs;
336        proc print data=mylib.dat(obs=10);
337        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              743.12k
      OS Memory           27048.00k
      Timestamp           05/01/2025 05:24:23 PM
      Step Count                        35  Switch Count  0
      Page Faults                       0
      Page Reclaims                     69
      Page Swaps                        0
      Voluntary Context Switches        9
      Involuntary Context Switches      3
      Block Input Operations            0
      Block Output Operations           16


338
339        * get frequency tables;
340        proc freq data=mylib.dat;
341        tables year month region type pres_party;
342        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       0.05 seconds
      system cpu time     0.00 seconds
      memory              1136.00k
      OS Memory           27308.00k
      Timestamp           05/01/2025 05:24:23 PM
      Step Count                        36  Switch Count  2
      Page Faults                       0
      Page Reclaims                     203
      Page Swaps                        0
      Voluntary Context Switches        24
      Involuntary Context Switches      4
      Block Input Operations            0
      Block Output Operations           280
      

343        
344        * describe dataset;
345        proc contents data=mylib.dat;
346        run;

NOTE: PROCEDURE CONTENTS used (Total process time):
      real time           0.03 seconds
      user cpu time       0.03 seconds
      system cpu time     0.01 seconds
      memory              1247.37k
      OS Memory           27308.00k
      Timestamp           05/01/2025 05:24:23 PM
      Step Count                        37  Switch Count  0
      Page Faults                       0
      Page Reclaims                     247
      Page Swaps                        0
      Voluntary Context Switches        7
      Involuntary Context Switches      2
      Block Input Operations            0
      Block Output Operations           24
      

347        
348        * END OF PROCESSING SCRIPT;
349        
350        ****************************************************************************
351        * Analysis Script
352        * Murphy John
353        * 2025-04-21
354        * This script performs the main analyses of this project.
355        ****************************************************************************;
356        
357        title "Setup";
358        
359        ** footnote;
360        footnote "Analysis script run on &SYSDATE at &SYSTIME.";
361        
362        ** establish library;
363        libname mylib "/home/u63984496/BIOS7400/final-project";
NOTE: Libref MYLIB was successfully assigned as follows: 
      Engine:        V9 
      Physical Name: /home/u63984496/BIOS7400/final-project
364        
365        ** set graphics ods;
366        ods graphics on / width=8in height=4in;
367        ****************************************************************************;
368        
369        title "Exploratory analysis";
370        
371        * plot the outcome of interest, avgprice;
372        ** sort data by type, date, and region;
373        proc sort data=mylib.dat;
374            by type date region;
375        run;

NOTE: There were 546 observations read from the data set MYLIB.DAT.
NOTE: The data set MYLIB.DAT has 546 observations and 18 variables.
NOTE: PROCEDURE SORT used (Total process time):
      real time           0.01 seconds
      user cpu time       0.00 seconds
      system cpu time     0.00 seconds
      memory              930.03k
      OS Memory           27308.00k
      Timestamp           05/01/2025 05:24:23 PM
      Step Count                        38  Switch Count  1
      Page Faults                       0
      Page Reclaims                     113
      Page Swaps                        0
      Voluntary Context Switches        40
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           272
      

376        
377        ** plot price over time by type stratified by region;
```

```
378        proc sgpanel data=mylib.dat;
379            panelby type;
380            series x = date y = avgprice / group = region;
381        run;

NOTE: PROCEDURE SGPANEL used (Total process time):
      real time           2.31 seconds
      user cpu time       0.05 seconds
      system cpu time     0.03 seconds
      memory              11434.15k
      OS Memory           37552.00k
      Timestamp           05/01/2025 05:24:25 PM
      Step Count                        39  Switch Count  10
      Page Faults                       0
      Page Reclaims                     3146
      Page Swaps                        0
      Voluntary Context Switches        3653
      Involuntary Context Switches      1
      Block Input Operations            0
      Block Output Operations           1656

NOTE: The column format YYMMDD10 is replaced by an auto-generated format on the axis.
NOTE: The column format YYMMDD10 is replaced by an auto-generated format on the axis.
NOTE: The column format YYMMDD10 is replaced by an auto-generated format on the axis.
NOTE: The column format YYMMDD10 is replaced by an auto-generated format on the axis.
NOTE: The column format YYMMDD10 is replaced by an auto-generated format on the axis.
NOTE: The column format YYMMDD10 is replaced by an auto-generated format on the axis.
NOTE: There were 546 observations read from the data set MYLIB.DAT.

382
383        ** plot price by pres_party and type stratified by region;
384        proc sgpanel data=mylib.dat;
385        panelby region type;
386            hbox avgprice / group = pres_party;
387        run;

NOTE: PROCEDURE SGPANEL used (Total process time):
      real time           1.19 seconds
      user cpu time       0.25 seconds
      system cpu time     0.10 seconds
      memory              3412.03k
      OS Memory           36532.00k
      Timestamp           05/01/2025 05:24:27 PM
      Step Count                        40  Switch Count  98
      Page Faults                       0
      Page Reclaims                     3403
      Page Swaps                        0
      Voluntary Context Switches        26647
      Involuntary Context Switches      24
      Block Input Operations            0
      Block Output Operations           6920

NOTE: There were 546 observations read from the data set MYLIB.DAT.

388
389        ** plot price by temp and type stratified by region;
390        proc sgpanel data=mylib.dat;
391        panelby type;
392            scatter x = temp y = avgprice / group = region;
393        run;

NOTE: PROCEDURE SGPANEL used (Total process time):
      real time           0.22 seconds
      user cpu time       0.04 seconds
      system cpu time     0.01 seconds
      memory              3155.68k
      OS Memory           36788.00k
      Timestamp           05/01/2025 05:24:27 PM
      Step Count                        41  Switch Count  10
      Page Faults                       0
      Page Reclaims                     590
      Page Swaps                        0
      Voluntary Context Switches        3486
      Involuntary Context Switches      2
      Block Input Operations            0
      Block Output Operations           1160

NOTE: There were 546 observations read from the data set MYLIB.DAT.

394        ****************************************************************************;
395
396        * univariable analysis;
397        ** get means for each type;
398        proc univariate data=mylib.dat plots;
```

```
399         var avgprice;
400         class type;
401         run;

NOTE: PROCEDURE UNIVARIATE used (Total process time):
      real time           0.27 seconds
      user cpu time       0.13 seconds
      system cpu time     0.01 seconds
      memory              3376.87k
      OS Memory           36108.00k
      Timestamp           05/01/2025 05:24:27 PM
      Step Count                        42  Switch Count  0
      Page Faults                       0
      Page Reclaims                     545
      Page Swaps                        0
      Voluntary Context Switches        396
      Involuntary Context Switches      4
      Block Input Operations            0
      Block Output Operations           648


402
403         * bivariable analysis;
404         ** check correlations of numerical variables;
405         proc corr pearson data=mylib.dat;
406         var temp avgprice;
407         by type;
408         run;

NOTE: PROCEDURE CORR used (Total process time):
      real time           0.04 seconds
      user cpu time       0.04 seconds
      system cpu time     0.00 seconds
      memory              1180.25k
      OS Memory           35240.00k
      Timestamp           05/01/2025 05:24:27 PM
      Step Count                        43  Switch Count  0
      Page Faults                       0
      Page Reclaims                     66
      Page Swaps                        0
      Voluntary Context Switches        7
      Involuntary Context Switches      1
      Block Input Operations            0
      Block Output Operations           16


409
410         ** create macro to check means of categorical variables;
411         %macro means(var);
412         proc means data = mylib.dat;
413         var avgprice;
414         class &var;
415         by type;
416         run;
417         %mend means;
418
419         * run;
420         %means(month_num);

NOTE: There were 546 observations read from the data set MYLIB.DAT.
NOTE: PROCEDURE MEANS used (Total process time):
      real time           0.03 seconds
      user cpu time       0.03 seconds
      system cpu time     0.00 seconds
      memory              2110.89k
      OS Memory           36524.00k
      Timestamp           05/01/2025 05:24:27 PM
      Step Count                        44  Switch Count  2
      Page Faults                       0
      Page Reclaims                     222
      Page Swaps                        0
      Voluntary Context Switches        15
      Involuntary Context Switches      2
      Block Input Operations            0
      Block Output Operations           0


421         %means(year);

NOTE: There were 546 observations read from the data set MYLIB.DAT.
NOTE: PROCEDURE MEANS used (Total process time):
      real time           0.02 seconds
      user cpu time       0.02 seconds
      system cpu time     0.00 seconds
      memory              2065.67k
      OS Memory           36524.00k
      Timestamp           05/01/2025 05:24:27 PM
      Step Count                        45  Switch Count  2
      Page Faults                       0
      Page Reclaims                     180
      Page Swaps                        0
      Voluntary Context Switches        15
      Involuntary Context Switches      1
      Block Input Operations            0
      Block Output Operations           16


422        %means(region);

NOTE: There were 546 observations read from the data set MYLIB.DAT.
NOTE: PROCEDURE MEANS used (Total process time):
      real time           0.02 seconds
      user cpu time       0.02 seconds
      system cpu time     0.01 seconds
      memory              2047.45k
      OS Memory           36524.00k
      Timestamp           05/01/2025 05:24:27 PM
      Step Count                        46  Switch Count  2
      Page Faults                       0
      Page Reclaims                     180
      Page Swaps                        0
      Voluntary Context Switches        14
      Involuntary Context Switches      2
      Block Input Operations            0
      Block Output Operations           16


423        %means(pres_party);

NOTE: There were 546 observations read from the data set MYLIB.DAT.
NOTE: PROCEDURE MEANS used (Total process time):
      real time           0.02 seconds
      user cpu time       0.02 seconds
      system cpu time     0.00 seconds
      memory              2026.23k
      OS Memory           36524.00k
      Timestamp           05/01/2025 05:24:27 PM
      Step Count                        47  Switch Count  2
      Page Faults                       0
      Page Reclaims                     180
      Page Swaps                        0
      Voluntary Context Switches        14
      Involuntary Context Switches      1
      Block Input Operations            0
      Block Output Operations           0


424        ****************************************************************************;
425
426        title "Regression Model fits";
427
428        * create train and test data;
429        ** randomly select 80 percent of the data for the training and reserve the remainder for testing;
430        proc surveyselect data=mylib.dat
431            out=dat_select
432            samprate=0.8
433            outall
434            seed=333;
435        run;

NOTE: The data set WORK.DAT_SELECT has 546 observations and 19 variables.
NOTE: PROCEDURE SURVEYSELECT used (Total process time):
      real time           0.01 seconds
      user cpu time       0.01 seconds
      system cpu time     0.00 seconds
      memory              919.25k
      OS Memory           35756.00k
      Timestamp           05/01/2025 05:24:27 PM
      Step Count                        48  Switch Count  2
      Page Faults                       0
      Page Reclaims                     249
      Page Swaps                        0
      Voluntary Context Switches        21
      Involuntary Context Switches      1
      Block Input Operations            0
      Block Output Operations           264
```

```

           ** create seperate data sets;
           data dat_train dat_test;
               set dat_select;
               if selected then output dat_train;
               else output dat_test;
           run;

NOTE: There were 546 observations read from the data set WORK.DAT_SELECT.
NOTE: The data set WORK.DAT_TRAIN has 437 observations and 19 variables.
NOTE: The data set WORK.DAT_TEST has 109 observations and 19 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              1324.31k
      OS Memory           36016.00k
      Timestamp           05/01/2025 05:24:27 PM
      Step Count                        49  Switch Count  4
      Page Faults                       0
      Page Reclaims                     157
      Page Swaps                        0
      Voluntary Context Switches        28
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           528




           * simple linear regression;
           ** avgprice by type, only;
           proc glm data=dat_train plots=all;
           class type (ref='conventional');
           model avgprice = type;
           run;


           *** rmse = 0.251153, rsq = 0.485679;

           * bivariable regression with interaction;
           ** model avgprice by type, covariate, and interaction;
           ** covariates are month_num, year, temp, region, pres_party;

           * write macro to fit the numerical covariate model;
           %macro bivariable_num(covariate);
           proc glm data=dat_train plots=all;
                   class type (ref='conventional');
                   model avgprice = type &covariate type*&covariate;
               run;

           %mend bivariable_num;

           * write macro to fit the categorical covariate models;
           %macro bivariable_cat(covariate);
           proc glm data=dat_train plots=all;
                   class type (ref='conventional') &covariate;
                   model avgprice = type &covariate type*&covariate;
               run;

           %mend bivariable_cat;

           * run;

           %bivariable_num(temp);

NOTE: PROCEDURE GLM used (Total process time):
      real time           0.43 seconds
      user cpu time       0.14 seconds
      system cpu time     0.02 seconds
      memory              10400.68k
      OS Memory           42972.00k
      Timestamp           05/01/2025 05:24:28 PM
      Step Count                        50  Switch Count  23
      Page Faults                       0
      Page Reclaims                     12000
      Page Swaps                        0
      Voluntary Context Switches        764
      Involuntary Context Switches      8
      Block Input Operations            0
      Block Output Operations           1480
```

```
479        *** rmse = 0.234790, rsq = 0.552579;
480
481        %bivariable_cat(region);

NOTE: PROCEDURE GLM used (Total process time):
      real time           0.47 seconds
      user cpu time       0.17 seconds
      system cpu time     0.03 seconds
      memory              10260.03k
      OS Memory           42972.00k
      Timestamp           05/01/2025 05:24:28 PM
      Step Count                        51  Switch Count  23
      Page Faults                       0
      Page Reclaims                     11836
      Page Swaps                        0
      Voluntary Context Switches        2132
      Involuntary Context Switches      8
      Block Input Operations            0
      Block Output Operations           2480



482        *** rmse = 0.202991, rsq = 0.673291;
483
484        %bivariable_cat(pres_party);

NOTE: PROCEDURE GLM used (Total process time):
      real time           0.41 seconds
      user cpu time       0.14 seconds
      system cpu time     0.03 seconds
      memory              10189.37k
      OS Memory           43228.00k
      Timestamp           05/01/2025 05:24:29 PM
      Step Count                        52  Switch Count  23
      Page Faults                       0
      Page Reclaims                     11613
      Page Swaps                        0
      Voluntary Context Switches        3429
      Involuntary Context Switches      6
      Block Input Operations            0
      Block Output Operations           1704



485        *** rmse = 0.244407, rsq = 0.515176;
486
487        %bivariable_cat(year);

NOTE: PROCEDURE GLM used (Total process time):
      real time           0.32 seconds
      user cpu time       0.13 seconds
      system cpu time     0.03 seconds
      memory              10154.53k
      OS Memory           43228.00k
      Timestamp           05/01/2025 05:24:29 PM
      Step Count                        53  Switch Count  23
      Page Faults                       0
      Page Reclaims                     11588
      Page Swaps                        0
      Voluntary Context Switches        3336
      Involuntary Context Switches      7
      Block Input Operations            0
      Block Output Operations           1536



488        *** rmse = 0.239643, rsq = 0.538200;
489
490        %bivariable_cat(month_num);

NOTE: PROCEDURE GLM used (Total process time):
      real time           0.33 seconds
      user cpu time       0.13 seconds
      system cpu time     0.03 seconds
      memory              10292.59k
      OS Memory           43484.00k
      Timestamp           05/01/2025 05:24:29 PM
      Step Count                        54  Switch Count  23
      Page Faults                       0
      Page Reclaims                     11569
      Page Swaps                        0
      Voluntary Context Switches        3369
      Involuntary Context Switches      6
      Block Input Operations            0
      Block Output Operations           1568



491        *** rmse = 0.231129, rsq = 0.586450;
492
493        * create table of metrics by hand;

NOTE: PROCEDURE GLM used (Total process time):
      real time           0.32 seconds
      user cpu time       0.15 seconds
      system cpu time     0.03 seconds
      memory              10310.34k
      OS Memory           43484.00k
      Timestamp           05/01/2025 05:24:30 PM
      Step Count                        55  Switch Count  23
      Page Faults                       0
      Page Reclaims                     11670
      Page Swaps                        0
      Voluntary Context Switches        3511
      Involuntary Context Switches      5
      Block Input Operations            0
      Block Output Operations           1760


494        data stats;
495            length Covariate $30;
496            infile datalines dsd truncover;
497            input Covariate :$30. R_Square RMSE;
498            datalines;

NOTE: The data set WORK.STATS has 6 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              675.37k
      OS Memory           36520.00k
      Timestamp           05/01/2025 05:24:30 PM
      Step Count                        56  Switch Count  2
      Page Faults                       0
      Page Reclaims                     123
      Page Swaps                        0
      Voluntary Context Switches        12
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           264


505        ;
506        run;
507
508        proc print data=stats noobs;
509        var Covariate RMSE R_Square;
510        run;

NOTE: There were 6 observations read from the data set WORK.STATS.
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              610.56k
      OS Memory           36520.00k
      Timestamp           05/01/2025 05:24:30 PM
      Step Count                        57  Switch Count  0
      Page Faults                       0
      Page Reclaims                     100
      Page Swaps                        0
      Voluntary Context Switches        0
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           0


511
512        *** All three bivariable models improve the fit above the univariable model as measued by the rmse and r-squared.
513        *** The bivariable model with region has the lowest rmse and highest r-squared.;
514
515        * full model;
516        ** avgprice by type, temp, region, pres_party and interactions;
517        proc glm data=dat_train plots=all;
518                class pres_party type (ref='conventional') region month_num year;
519                model avgprice = type temp region pres_party month_num year
520                  type*temp type*region type*pres_party type*month_num type*year;
```

```
521        run;

522        *** rmse = 0.145137, rsq = 0.844828;
523        
524        * reduced model;
525        ** the interaction between type and temp, month_num, and pres_party have small Type III SS and large p-values.
526        ** Reduce the model by removing these terms.;

NOTE: PROCEDURE GLM used (Total process time):
      real time           0.28 seconds
      user cpu time       0.14 seconds
      system cpu time     0.02 seconds
      memory              10340.18k
      OS Memory           43228.00k
      Timestamp           05/01/2025 05:24:30 PM
      Step Count                        58  Switch Count  23
      Page Faults                       0
      Page Reclaims                     11694
      Page Swaps                        0
      Voluntary Context Switches        607
      Involuntary Context Switches      7
      Block Input Operations            0
      Block Output Operations           1784
      

527        proc glm data=dat_train plots=all alpha=0.05;
528                class type (ref='conventional') region pres_party month_num year;
529                model avgprice = type temp region pres_party month_num year
530                  type*region;
531                store out=final_model;
532        run;

533        quit;

NOTE: The GLM procedure generated the model item store WORK.FINAL_MODEL.
NOTE: PROCEDURE GLM used (Total process time):
      real time           0.29 seconds
      user cpu time       0.14 seconds
      system cpu time     0.02 seconds
      memory              10623.87k
      OS Memory           43228.00k
      Timestamp           05/01/2025 05:24:30 PM
      Step Count                        59  Switch Count  23
      Page Faults                       0
      Page Reclaims                     11698
      Page Swaps                        0
      Voluntary Context Switches        602
      Involuntary Context Switches      7
      Block Input Operations            0
      Block Output Operations           2040
      

534        *** rmse = 0.152099, rsq = 0.823078;
535        
536        *** The reduced model is very similar in terms of fit to the full model but does not include unnessecary interaction
536      ! terms.
537        *** Select this as the final model.;
538        ****************************************************************************;
539        
540        title 'Model Evaluation';
541        * use the final model to generate predictions on the test data;
542        proc plm restore=final_model;
543            score data=dat_test out=predictions predicted;
544        run;

NOTE: The data set WORK.PREDICTIONS has 109 observations and 20 variables.
NOTE: PROCEDURE PLM used (Total process time):
      real time           0.01 seconds
      user cpu time       0.02 seconds
      system cpu time     0.00 seconds
      memory              1136.65k
      OS Memory           36524.00k
      Timestamp           05/01/2025 05:24:30 PM
      Step Count                        60  Switch Count  2
      Page Faults                       0
      Page Reclaims                     156
      Page Swaps                        0
      Voluntary Context Switches        14
      Involuntary Context Switches      1
      Block Input Operations            0
      Block Output Operations           272
      

545        
```

```
           * compute rmse and r-squared;
           ** get residuals;
           data eval;
               set predictions;
               resid = avgprice - predicted;
               sq_resid = resid**2;
           run;

NOTE: There were 109 observations read from the data set WORK.PREDICTIONS.
NOTE: The data set WORK.EVAL has 109 observations and 22 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              986.62k
      OS Memory           36524.00k
      Timestamp           05/01/2025 05:24:30 PM
      Step Count                        61  Switch Count  2
      Page Faults                       0
      Page Reclaims                     114
      Page Swaps                        0
      Voluntary Context Switches        14
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           272



           * compute mean;
           proc means data=eval noprint;
               var avgprice predicted sq_resid;
               output out=metrics
                   mean(avgprice)=mean_y
                   sum(sq_resid)=ss_res
                   n=samples;
           run;

NOTE: There were 109 observations read from the data set WORK.EVAL.
NOTE: The data set WORK.METRICS has 1 observations and 5 variables.
NOTE: PROCEDURE MEANS used (Total process time):
      real time           0.00 seconds
      user cpu time       0.00 seconds
      system cpu time     0.01 seconds
      memory              6646.78k
      OS Memory           42432.00k
      Timestamp           05/01/2025 05:24:30 PM
      Step Count                        62  Switch Count  3
      Page Faults                       0
      Page Reclaims                     1538
      Page Swaps                        0
      Voluntary Context Switches        32
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           264



           * compute metrics;
           data results;
               set metrics;
               ss_total = 0;
               do i = 1 to samples;
                   set eval point=i nobs=n;
                   ss_total + (avgprice - mean_y)**2;
               end;
               rmse = sqrt(ss_res / samples);
               rsq = 1 - (ss_res / ss_total);
               keep rmse rsq;
           run;

NOTE: There were 1 observations read from the data set WORK.METRICS.
NOTE: The data set WORK.RESULTS has 1 observations and 2 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              1338.75k
      OS Memory           36528.00k
      Timestamp           05/01/2025 05:24:30 PM
      Step Count                        63  Switch Count  3
      Page Faults                       0
      Page Reclaims                     154
      Page Swaps                        0
      Voluntary Context Switches        17
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           264


575
576        proc print data=results;
577            title "RMSE and R-squared on Test Data";
578        run;

NOTE: There were 1 observations read from the data set WORK.RESULTS.
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              616.15k
      OS Memory           36008.00k
      Timestamp           05/01/2025 05:24:30 PM
      Step Count                        64  Switch Count  1
      Page Faults                       0
      Page Reclaims                     62
      Page Swaps                        0
      Voluntary Context Switches        8
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           0


579
580        * residuals vs fitted plot;
581        proc sgplot data=eval;
582            scatter x=predicted y=resid / markerattrs=(symbol=circlefilled color=black);
583            refline 0 / axis=y lineattrs=(color=red pattern=shortdash);
584            xaxis label="Fitted Values";
585            yaxis label="Residuals";
586            title "Residuals vs. Fitted Values";
587        run;

NOTE: PROCEDURE SGPLOT used (Total process time):
      real time           0.07 seconds
      user cpu time       0.02 seconds
      system cpu time     0.00 seconds
      memory              1656.46k
      OS Memory           36876.00k
      Timestamp           05/01/2025 05:24:30 PM
      Step Count                        65  Switch Count  2
      Page Faults                       0
      Page Reclaims                     299
      Page Swaps                        0
      Voluntary Context Switches        142
      Involuntary Context Switches      3
      Block Input Operations            0
      Block Output Operations           384

NOTE: There were 109 observations read from the data set WORK.EVAL.

588
589        * predicted vs observed plot;
590        proc sgplot data=predictions;
591            scatter x=predicted y=avgprice / markerattrs=(symbol=circlefilled color=black);
592            lineparm x=0 y=0 slope=1 / lineattrs=(color=red pattern=shortdash);
593            xaxis label="Predicted Values";
594            yaxis label="Observed Values";
595            title "Predicted vs. Observed Values";
596        run;

NOTE: PROCEDURE SGPLOT used (Total process time):
      real time           0.07 seconds
      user cpu time       0.02 seconds
      system cpu time     0.00 seconds
      memory              2177.71k
      OS Memory           36876.00k
      Timestamp           05/01/2025 05:24:30 PM
      Step Count                        66  Switch Count  2
      Page Faults                       0
      Page Reclaims                     292
      Page Swaps                        0
      Voluntary Context Switches        141
      Involuntary Context Switches      1
      Block Input Operations            0
      Block Output Operations           408

NOTE: There were 109 observations read from the data set WORK.PREDICTIONS.
```

```

* END OF ANALYSIS SCRIPT;

* END OF AGGREGATED SCRIPT;

OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;

```