```
*************************************************************************
* Process Data Script
* Murphy John
* 2025-04-07
* This script loads, processes, and compiles the data used in this project.
*************************************************************************;

title "Setup";

** footnote;
footnote "Data processing script run on &SYSDATE at &SYSTIME.";

** establish library;
libname mylib "/home/u63984496/BIOS7400/final-project";
*************************************************************************;

title "Data processing";

title2 "Avocado Data";

* load data;
proc import datafile="/home/u63984496/BIOS7400/final-project/avocado.csv"
    out=work.raw_avo
    dbms=csv
    replace;
    guessingrows=MAX;
run;

* 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
        ));

    * seperate date by month and year;
    * create a month year variable;
    month = put(date, monname.);
    month_num = month(date);
    month = strip(propcase(month));
    date = mdy(month_num, 1, year);

    * keep only specififc regions;
    if region not in (
        "California",
        "West",
        "Northeast",
        "SouthCentral",
        "Southeast",
        "GreatLakes",
        "MidSouth",
        "Plains")
        then delete;

    drop VAR1;
run;
```

```
* group by year, month, region, and type;
proc sql;
    create table work.avo_group as
    select
        year,
        month,
        month_num,
        date,
        region,
        type,
        mean(avgprice) as avgprice format=8.2,
        sum(totvol) as totvol,
        sum(totsm) as totsm,
        sum(totlg) as totlg,
        sum(totxl) as totxl,
        sum(totbags) as totbags,
        sum(totbags_sm) as totbags_sm,
        sum(totbags_lg) as totbags_lg,
        sum(totbags_xl) as totbags_xl
    from work.clean_avo
    group by date, region, type;
quit;

* sort by date and remove duplicate obs;
proc sort data=work.avo_group nodupkey out=work.dat_avo;
    by date region type;
run;

* print first 10 obs;
proc print data=work.dat_avo(obs=10);
run;

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;
    keep year month temp
```

```
run;

* print first 10 obs;
proc print data=work.dat_temp(obs=10);
run;

title2 "President Data";

*** In 2015 and 2016, Barack Obama of the democratic party was president of the US.
*** In 2017 and 2018, Donald Trump of the republican party was president of the US.;

* establish data;
data dat_pres;
    length year 4 president $ 20 pres_party $ 25;
    input year president pres_party;
    infile datalines dsd dlm = " ";
    datalines;
2015 "Barack Obama" "Democratic"
2016 "Barack Obama" "Democratic"
2017 "Donald Trump" "Republican"
2018 "Donald Trump" "Republican"
;
run;

* print;
proc print data=work.dat_pres;
run;
*******************************************************************************;

title "Data merging";
* sql can handle many-to-one merging;
* save to mylib;
proc sql;
    create table work.dat_merge as
    select
        a.*,
        b.*,
        c.*
    from work.dat_avo as a
    inner join work.dat_temp as b
        on a.year = b.year and a.month = b.month
    inner join work.dat_pres as c
        on a.year = c.year;
quit;

* add labels to variables;
data mylib.dat;
    set work.dat_merge;
    label
        year = "Year"
        month = "Month Name"
        month_num = "Month Number"
        date = "Date of observation- only month and years are known"
        region = "City or region of the observation"
        type = "Type of farming method"
        avgprice = "Average price of a single avocado"
        totvol = "Total Number of avocados sold"
        totsm = "Total number of avocados with PLU 4046 (small) sold"
        totlg = "Total number of avocados with PLU 4225 (large) sold"
        totxl = "Total number of avocados with PLU 4770 (xlarge) sold"
        totbags = "Total number of bags sold"
        totbags_sm = "Total number of PLU 4046 (small) bags sold"
        totbags_lg = "Total number of PLU 4225 (large) bags sold"
        totbags_xl = "Total number of PLU 4770 (xlarge) bags sold"
        temp = "Temperature difference (degress C)"
        president = "Name of current U.S. president"
```

```
        pres_party = "Poliical Party of current U.S. president";
run;
*************************************************************************;

title "Print data";

* print first 10 obs;
proc print data=mylib.dat(obs=10);
run;

* get frequency tables;
proc freq data=mylib.dat;
    tables year month region type pres_party;
run;

* describe dataset;
proc contents data=mylib.dat;
run;
```