Programming Exercises

23. *Hint:* You can create the histograms for all the continents in one SGPLOT, and similarly for the box plots.

*Answer:*

LIBNAME sasdata 'c:\MySASLib';

\*\* Part a);

PROC SGPLOT DATA = sasdata.population;

HISTOGRAM Y1;

TITLE 'International Population';

RUN;

\*\* Part b);

PROC SORT DATA = sasdata.population OUT = popsort;

BY continent;

RUN;

PROC SGPLOT DATA = popsort;

HISTOGRAM Y1;

BY continent;

TITLE 'International Population by Continent';

RUN;

\*\* Part c);

PROC SGPLOT DATA = popsort;

HBOX Y1 / CATEGORY = continent;

TITLE 'International Population by Continent';

RUN;

\*\* Part d);

\*\* Both the histograms and box plots show the shape of

the data. The box plot also shows the minimum,

maximum, mean, median, and interquartile range as

well as markers for observations that are considered

outliers;

(sections 8.3, 8.4)

24. *Hint:* Create a customized format to label missing values. Create a new data set for the paneled graph so that there is one observation for each survey question for each subject.

*Answer:*

LIBNAME sasdata 'c:\MySASLib';

\*\* Part a);

PROC FORMAT;

VALUE resp . = 'Missing';

RUN;

PROC SGPLOT DATA = sasdata.wls;

VBAR Q1 / MISSING;

FORMAT Q1 resp.;

TITLE 'Responses to Question 1';

RUN;

PROC SGPLOT DATA = sasdata.wls;

\*\* Parts b) and c);

VBAR Q1 / MISSING RESPONSE = BMI

STAT = MEAN LIMITSTAT = STDERR;

\*\* Part d);

LABEL BMI = 'Mean Body Mass Index';

YAXIS VALUES = (15 TO 30 BY 5);

FORMAT Q1 resp.;

TITLE 'Average BMI for Question 1';

RUN;

\*\* Part e);

DATA wlst (KEEP = ID BMI SurvQues Response);

SET sasdata.wls;

ARRAY q (30) Q1 - Q30;

DO SurvQues = 1 TO 30;

Response = q(SurvQues);

OUTPUT;

END;

LABEL SurvQues = 'Survey Question';

RUN;

PROC SGPANEL DATA = wlst;

PANELBY SurvQues / COLUMNS = 3 ROWS = 2;

VBAR Response / MISSING RESPONSE = BMI

STAT = MEAN LIMITSTAT = STDERR;

LABEL BMI = 'Mean Body Mass Index';

ROWAXIS VALUES = (15 TO 30 BY 5);

FORMAT Response resp.;

TITLE 'Average BMI for Survey Questions';

RUN;

\*\* Part f);

ODS LISTING GPATH = 'c:\MyGraphics'

STYLE = ANALYSIS;

ODS GRAPHICS / RESET IMAGENAME = 'WLSgraph'

OUTPUTFMT = JPEG;

PROC SGPANEL DATA = wlst;

WHERE SurvQues IN (4,6,12,18);

PANELBY SurvQues;

VBAR Response / MISSING RESPONSE = BMI

STAT = MEAN LIMITSTAT = STDERR;

LABEL BMI = 'Mean Body Mass Index';

ROWAXIS VALUES = (15 TO 30 BY 5);

FORMAT Response resp.;

TITLE 'Average BMI for Survey Questions';

RUN;

(sections 8.1, 8.2, 8.8, 8.11, 8.12)

25. *Hint:* Use a DATA step to create a variable that can be plotted on the graph to identify the county with the largest number of patents.

*Answer:*

LIBNAME sasdata 'c:\MySASLib';

\*\* Part a);

PROC SGPLOT DATA = sasdata.patents;

WHERE Patents >= 100;

HISTOGRAM Patents;

DENSITY Patents;

TITLE 'Histogram of Patents';

RUN;

\*\* Part b);

\*\* The shape of this histogram is not normally

distributed, it appears to be skewed right;

\*\* Part c);

PROC SGPLOT DATA = sasdata.patents;

WHERE patents >= 100;

SCATTER X = education Y = patents;

TITLE 'Education Level by Patents';

RUN;

\*\* Part d);

PROC SORT DATA = sasdata.patents OUT = patsort;

BY DESCENDING Patents;

RUN;

\*\* Creating a new variable that contains the county name

for only the first observation from the sorted data

set will allow you to label only this county on the

plot;

DATA patlabel;

SET patsort;

IF \_N\_ = 1 THEN CntyLabel = County;

RUN;

PROC SGPLOT DATA = patlabel;

WHERE patents >= 100;

\*\* Parts d) and e);

SCATTER X = education Y = patents /

DATALABEL = CntyLabel

MARKERATTRS = (SYMBOL = CIRCLEFILLED)

TRANSPARENCY = 0.5;

TITLE 'Education Level by Patents';

RUN;

(sections 8.3, 8.5, 8.10)

26. *Hint:* This exercise will produce only one graphic. Create a variable for average magnitude that is combined with the original data set.

*Answer:*

LIBNAME sasdata 'c:\MySASLib';

\*\* Part b);

PROC SORT DATA = sasdata.earthquakes OUT = eqsort;

BY Year;

RUN;

PROC MEANS DATA = eqsort NOPRINT;

VAR Magnitude;

BY Year;

OUTPUT OUT = avemag MEAN(Magnitude) = EQmean;

RUN;

DATA avequake;

MERGE eqsort

avemag;

BY Year;

RUN;

PROC SGPLOT DATA = avequake;

\*\* Part a);

WHERE Year >= 2000;

SCATTER X = Year Y = Magnitude;

\*\* Part b);

SERIES X = Year Y = EQmean /

LINEATTRS = (COLOR = RED);

\*\* Part c);

KEYLEGEND / NOBORDER POSITION = BOTTOMRIGHT;

LABEL EQmean = 'Mean';

\*\* Part d);

REFLINE 4 5 6 7 8 /

LABEL = ('Light' 'Moderate' 'Strong'

'Major' 'Great')

LINEATTRS = (PATTERN = SHORTDASH)

TRANSPARENCY = 0.5;

\*\* Part e);

XAXIS TYPE = DISCRETE;

TITLE 'Magnitude of Earthquakes Since 2000';

RUN;

(sections 8.5, 8.6, 8.8, 8.9, 8.10)

27. *Hint:* For part c), create two regression plots with one REG statement by selecting the appropriate option.

*Answer:*

LIBNAME sasdata 'c:\MySASLib';

\*\* Part a);

PROC SGPLOT DATA = sasdata.study\_gpa;

HBOX AveTime / CATEGORY = Section;

TITLE 'Time Studied by Section';

RUN;

\*\* Part b);

PROC SGPLOT DATA = sasdata.study\_gpa NOAUTOLEGEND;

REG X = AveTime Y = GPA;

TITLE 'Regression Plot of Study Time and GPA';

RUN;

\*\* Parts c) and d);

PROC SGPLOT DATA = sasdata.study\_gpa;

REG X = AveTime Y = GPA / GROUP = Section CLM

CLMTRANSPARENCY = 0.50;

KEYLEGEND / LOCATION = OUTSIDE POSITION = RIGHT;

TITLE1 'Regression Plot of Study Time and GPA';

TITLE2 'By Section';

RUN;

\*\* Part e);

\*\* The box plots show that section 01 has a higher

average study time than section 02;

\*\* The first regression plot suggests no linear

relationship between study time and GPA.

\*\* The second regression plot indicates that there is no

linear relationship for section 02, but there is a

potential decreasing linear relationship for section

01;

(sections 8.4, 8.7, 8.9)

28. *Hint:* The data will need some summarization before being plotted. To insert text inside the plot area for part e), you will need to add some new variables to the data set for the text and the placement of the text on the plot. These new variables will have missing values for all but two observations, and they will need to be specified in an additional plot statement.

*Answer:*

LIBNAME sasdata 'c:\MySASLib';

\*\* Part d);

PROC FORMAT;

VALUE Mil 0 = '0'

20000000 = '20'

40000000 = '40'

60000000 = '60'

80000000 = '80';

VALUE $Size 'L' = 'Large Hub'

'M' = 'Medium Hub'

'S' = 'Small';

RUN;

\*\* Part a);

PROC SORT DATA = sasdata.airlines OUT = airlines;

BY Airline Year AirportSize;

RUN;

PROC MEANS DATA = airlines NOPRINT;

BY Airline Year AirportSize;

VAR Passengers;

OUTPUT OUT = airlinebyyear SUM(Passengers) = Psum;

RUN;

\*\* Part b);

PROC SGPLOT DATA = airlinebyyear;

BY Airline;

SERIES X = Year Y = Psum / GROUP = AirportSize;

TITLE 'Passengers By Airline in the United States';

RUN;

\*\* Part c);

PROC SGPANEL DATA = airlinebyyear;

PANELBY Airline / NOVARNAME COLUMNS = 3 ROWS = 2;

SERIES X = Year Y = Psum / GROUP = AirportSize;

\*\* Part d);

LABEL Psum = 'Passengers (Millions)'

AirportSize = 'Airport Size';

FORMAT Psum mil. AirportSize $Size.;

TITLE 'Passengers By Airline in the United States';

RUN;

\*\* Part e);

DATA all;

SET airlinebyyear;

BY Airline;

IF FIRST.Airline THEN DO;

IF Airline = 'US Airways Inc.' THEN DO;

Yaxis = 75000000;

Note = '+ American West in 2006';

END;

ELSE IF Airline = 'Delta Air Lines Inc.' THEN DO;

Yaxis = 75000000;

Note = '+ Northwest in 2010';

END;

END;

RUN;

PROC SGPANEL DATA = all;

PANELBY Airline / NOVARNAME COLUMNS = 3 ROWS = 2;

SCATTER X = Year Y = Yaxis / DATALABEL = Note

MARKERATTRS = (COLOR = WHITE);

SERIES X = Year Y = Psum / GROUP = AirportSize;

LABEL Psum = 'Passengers (Millions)'

Yaxis = 'Passengers (Millions)'

AirportSize = 'Airport Size';

FORMAT Psum Yaxis mil. AirportSize $Size.;

TITLE 'Passengers By Airline in the United States';

RUN;

(sections 8.5, 8.6, 8.10, 8.11)