**Exercise 1:** Write a SAS program to do the following. Give the SAS instructions and output. Use the Air-Pollution data on p.39 (Table 1.5).

1. Import the Air-Pollution data set into SAS using an *infile* statement in a *data* *step*.

libname pw "H:\Werk\Multivariate Statistical Analysis\libs";

**data** pw.air\_pol;

infile "H:\Werk\Multivariate Statistical Analysis\Data\T1-5.dat" dlm=' ';

input x1 x2 x3 x4 x5 x6 x7;

**run**;

1. Print the data set in SAS using *proc print*. Check if the data was correctly imported.

**proc** **print** data=pw.air\_pol;

var x1 x2 x3 x4 x5 x6 x7;

**run**;

|  |
| --- |
| The SAS System |

| **Obs** | **x1** | **x2** | **x3** | **x4** | **x5** | **x6** | **x7** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **1** | 8 | 98 | 7 | 2 | 12 | 8 | 2 |
| **2** | 7 | 107 | 4 | 3 | 9 | 5 | 3 |
| **3** | 7 | 103 | 4 | 3 | 5 | 6 | 3 |
| **4** | 10 | 88 | 5 | 2 | 8 | 15 | 4 |
| **5** | 6 | 91 | 4 | 2 | 8 | 10 | 3 |
| **6** | 8 | 90 | 5 | 2 | 12 | 12 | 4 |
| **7** | 9 | 84 | 7 | 4 | 12 | 15 | 5 |
| **8** | 5 | 72 | 6 | 4 | 21 | 14 | 4 |
| **9** | 7 | 82 | 5 | 1 | 11 | 11 | 3 |
| **…** | … | … | … | … | ... | … | … |
| **37** | 5 | 86 | 3 | 1 | 6 | 12 | 2 |
| **38** | 5 | 86 | 7 | 2 | 13 | 18 | 2 |
| **39** | 7 | 79 | 7 | 4 | 9 | 25 | 3 |
| **40** | 7 | 79 | 5 | 2 | 8 | 6 | 2 |
| **41** | 6 | 68 | 6 | 2 | 11 | 14 | 3 |
| **42** | 8 | 40 | 4 | 3 | 6 | 5 | 2 |

1. Split this data set into two data sets (call them *Data1* and *Data2*).
2. *Data1* should contain only observations where variable HC=3.
3. *Data2* should contain the rest of the observations.

**data** pw.Data1 pw.Data2;

set pw.air\_pol ;

if x7 = **3** then output pw.Data1;

else output pw.Data2;

**run**;

1. Print *Data1* and *Data2* to check if the split was correctly done.

**proc** **print** data=pw.Data1;

var x1 x2 x3 x4 x5 x6 x7;

**run**;



**proc** **print** data=pw.Data2;

var x1 x2 x3 x4 x5 x6 x7;

**run**;



1. In a second data step, split *Data1* into two further data sets (call them *Above50* and *Below50*).
2. *Above50* should contain only observations where variable Solar radiation is >=50.
3. *Below50* should contain only observations where variable Solar radiation is <50.

**data** pw.Above50 pw.Below50;

set pw.air\_pol ;

if x2 >= **50** then output pw.Above50;

else if x2 < **50** then output pw.Below50;

**run**;

1. Calculate the summary statistics of the *Above50* data set using *proc means*. Use only the four variables CO, NO, NO2 and O3 for this step.

**proc** **means** data=pw.Above50;

var x3 x4 x5 x6;

**run**;

1. Lastly, export the:
2. *Above50* data set in step 5 to a word document (give the totals of the variables as well).
3. Output of the *proc means* in step 6 to the same word document.

ods listing close;

ods rtf file=" H:\Werk\Multivariate Statistical Analysis\libs \assignment\_2\_ex1.rtf";

**proc** **print** data=pw.above50;

var x1 x2 x3 x4 x5 x6 x7;

sum x1 x2 x3 x4 x5 x6 x7;

**run**;

**proc** **means** data=pw.Above50;

title 'Exercise 1';

var x3 x4 x5 x6;

**run**;

ods \_all\_ close;

ods listing;

title;

**Exercise 2:** Write a SAS program to do the following. Give the SAS instructions and output. Use the sales.csv on the P-drive.

1. Import the *sales.csv* data set into SAS using an *infile* statement in a *data* *step*. The data set contains 9 variables and 165 observations. This is a comma delimited file. Give the following labels to the variables:

*Employee\_ID, First\_Name, Surname, Gender, Salary, Job\_Title, Country, StartDate* and *EndDate.*

* Give the name *CompanySales* to the data in SAS.
* Use the following instruction to set the length of Job\_Title

length Job\_Title $ **20**;

* Use the following instruction in the data step with *StartDate* and *EndDate*.

input StartDate: date9. EndDate: ddmmyy10.

format StartDate EndDate mmddyy10.

* Create a variable: Employed=year(EndDate)-year(StartDate)

**data** pw.Company;

length Job\_Title $ **20**;

infile "\\sunrga.stb.sun.ac.za\home\unlisted\15338673\Documents\Multivariate Statistical Analysis 745\Data\sales.csv" dlm=',' firstobs=**1**;

input Employee\_ID $ First\_Name $ Surname $ Gender $ Salary $ Job\_Title $ Country $ StartDate: date9. EndDate: ddmmyy10.;

format StartDate EndDate mmddyy10.;

Employed=year(EndDate)-year(StartDate);

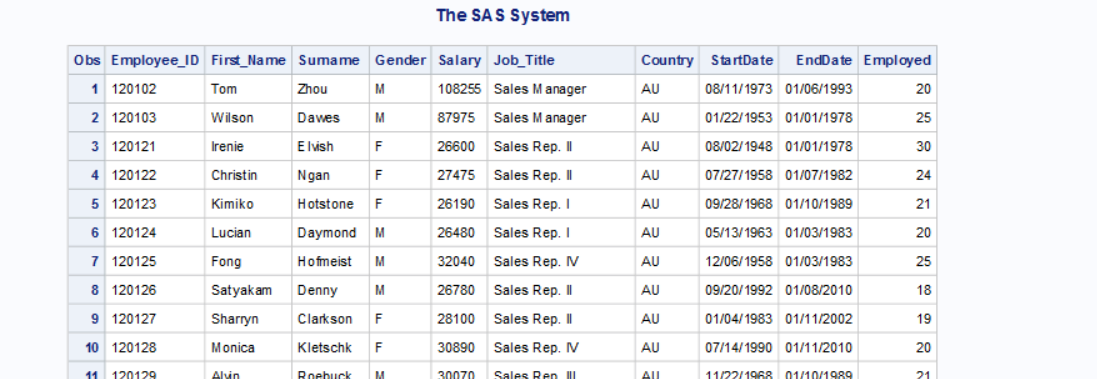
**run**;

1. Print the data set (*CompanySales*) in SAS using *proc print*. Check if the data was correctly imported.

**proc** **print** data=pw.Company;

var Employee\_ID First\_Name Surname Gender Salary Job\_Title Country StartDate EndDate Employed;

**run**;

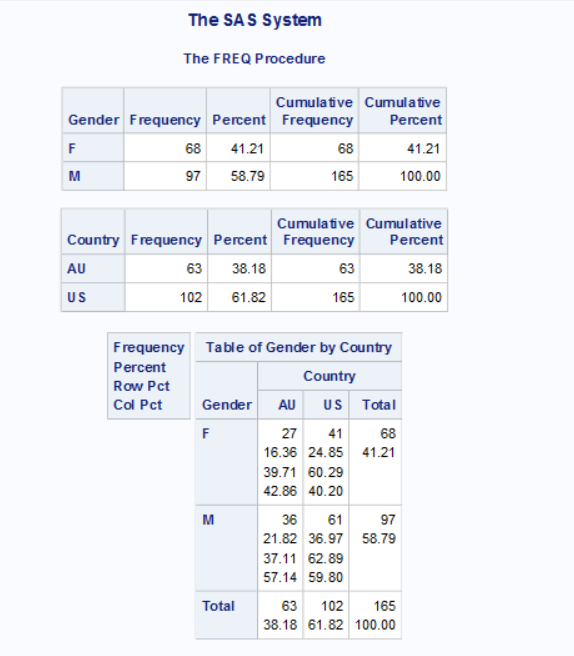


1. Use the *proc freq* to make the following:
2. A frequency table for Gender.
3. A frequency table for country.
4. A contingency table for Gender and Country.

**proc** **freq** data=pw.Company;

tables Gender Country Gender\*Country;

**run**;



1. Create two data sets (call them *RepData* and *USData*).
2. *RepData* should contain all the observations where *Rep* is part of the Job\_Title.
3. *USData* should contain all the observations for the country *US*.

**data** pw.RepData;

set pw.Company ;

if index(Job\_Title, 'Rep') > **0** then output pw.RepData;

**run**;

**data** pw.USData;

set pw.Company ;

if Country = 'US' then output pw.USData;

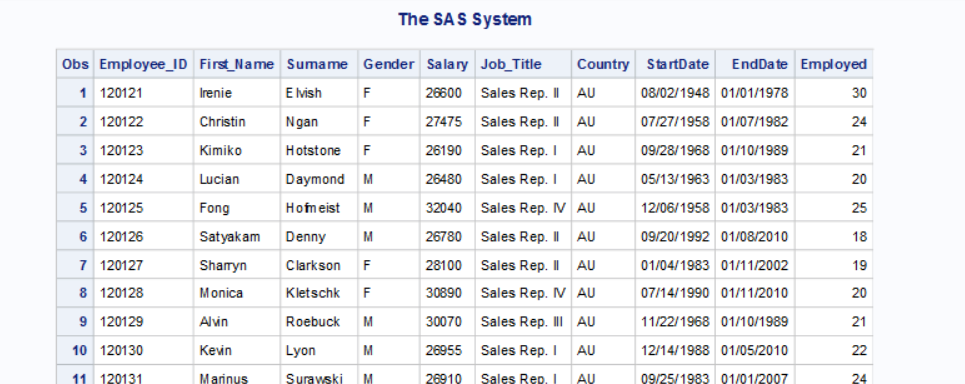
**run**;

1. Print *RepData* and *USData* to check if the data sets were create correctly.

**proc** **print** data=pw.RepData;

var Employee\_ID First\_Name Surname Gender Salary Job\_Title Country StartDate EndDate Employed;

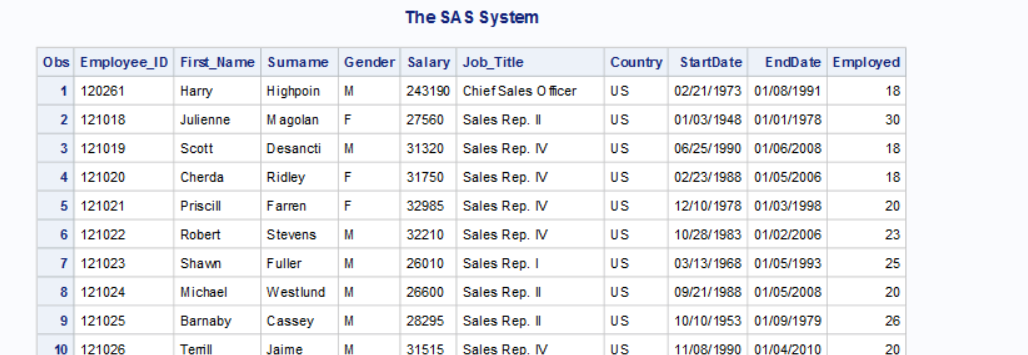
**run**;



**proc** **print** data=pw.USData;

var Employee\_ID First\_Name Surname Gender Salary Job\_Title Country StartDate EndDate Employed;

**run**;



1. In the *USData*, get only the First names and Surnames of employees who are female. Print the output.

**data** pw.USDataFemale;

set pw.USData ;

input First\_Name Surname;

if Gender = 'Female' then output pw.USDataFemale;

**run**;

**proc** **print** data=pw.USDataFemale;

var First\_Name Surname;

**run**;



1. In the *RepData*, get the only the First names, Surnames and Gender of the employees who are employed more than 20 years. Print your output.

**data** pw.RepDataLongEmploy;

set pw.USData ;

input First\_Name Surname Gender;

if Employed > **20** then output pw.RepDataLongEmploy;

**run**;

**proc** **print** data=pw.RepDataLongEmploy;

var First\_Name Surname Gender;

**run**;

