

**Department of Statistics and Actuarial Science**

**Multivariate Statistics**

**SAS Tutorial**

**by**

**Morné Lamont**

**All the data sets for this tutorial are stored under the directory**

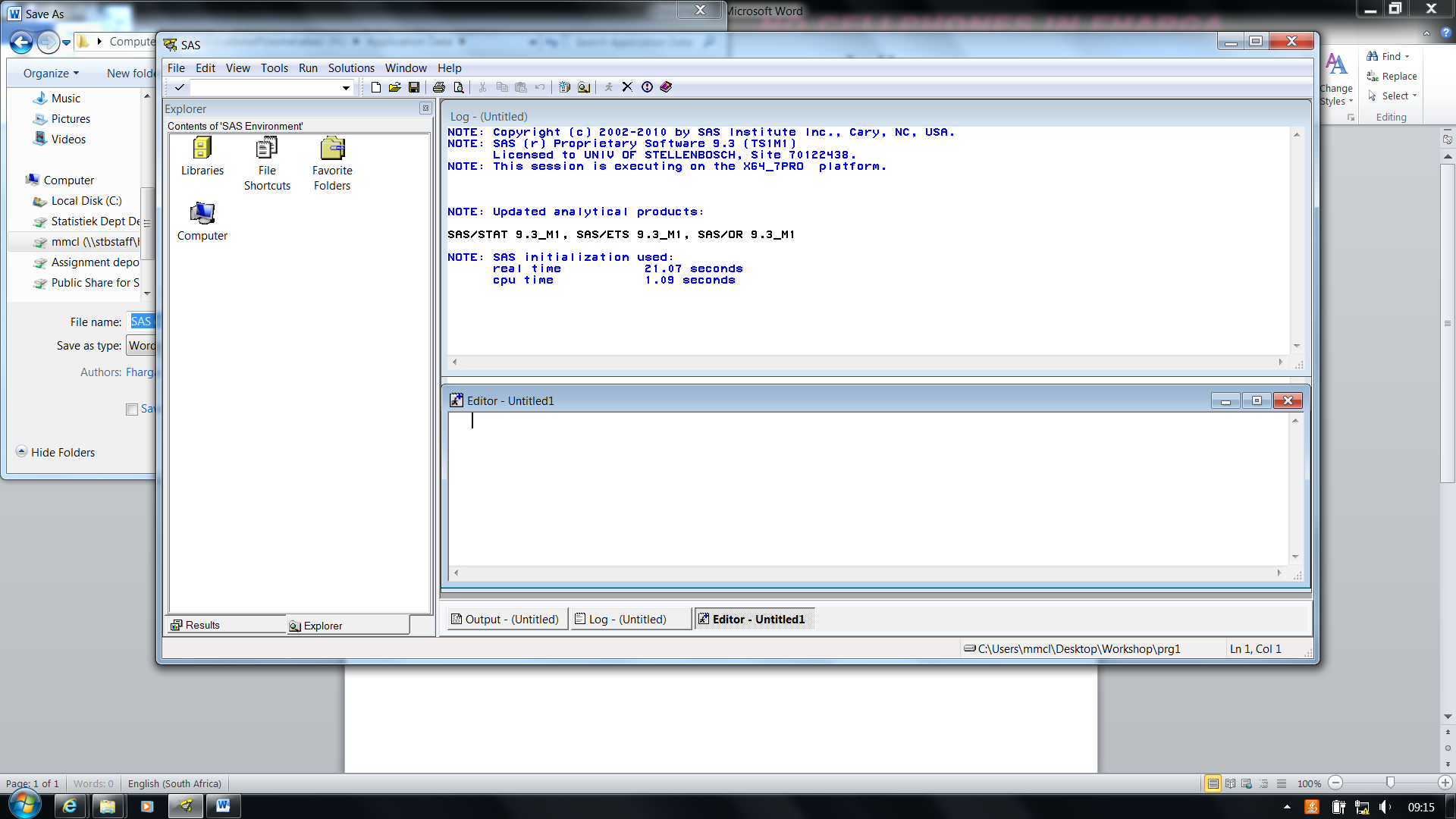
P:\ebw\2019\Statistiek\Nagraadse Modules\MVAR(10600721)\Data

**Data set names**

For **Lesson 1** we will use the following data sets: *Bone.xls*, *Bone.txt, BoneLabels.prn, Region.txt, AirPollution.txt and Sales.csv.*

For **Lesson 2 and 3** we will use the following data sets: *IrisData.txt, HimophiliaLearn.txt, HimophiliaTest.txt,*

**Setting the default path and creating a new library**



* Change the path to your preferred directory by double left clicking on the directory circled above.
* This is where you would put all your SAS data sets.

Type the following command in the editor widow.

/\* Creating a path to my H: directory \*/

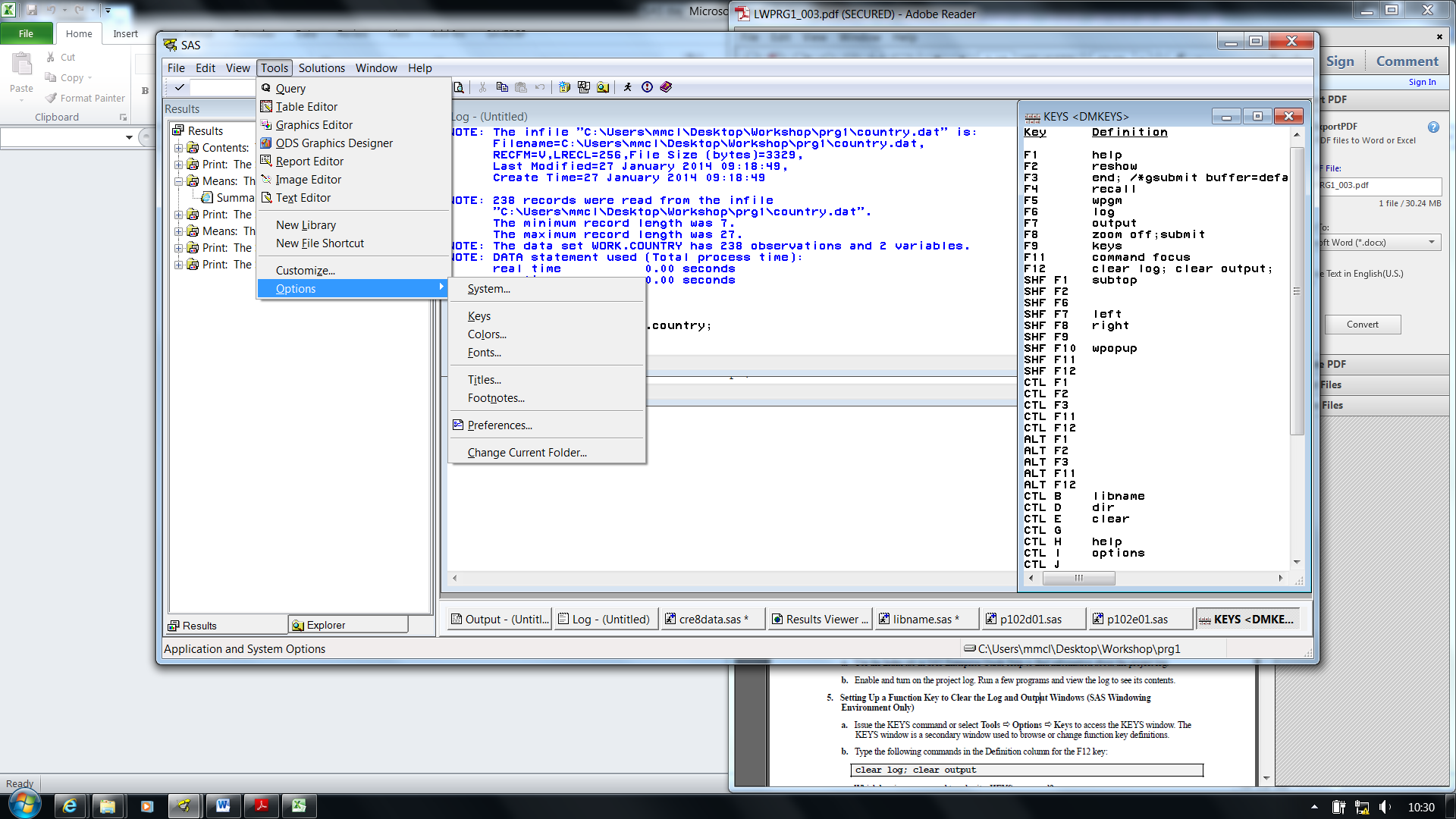
%let path=H:\Honours\SAS\_tutorial;

Create your own library under Explorer, by typing the following command in the editor window.

/\* Creating a library name called by your name ‘morne’ \*/

libname morne "H:\Honours\SAS\_tutorial";

**Using the keys and preferences options**



In **Tools - Options - Keys** you view the shortcuts.

* Use the **F1** key to go to the help file of SAS.
* Create shortcuts for yourself:

for example – in **F12** type the following **clear log; clear output;** The F12 key can now be used to clear the log and output window.

In **Tools - Options - Preferences** you can set the results format.

* Go to the **results** tab and select **Create listing** and **Create** **HTML**.
* Deselect **Use WORK folder**.
* Change the **Style**.

**Lesson 1: Data management**

**Scenario 1:**

**Importing a txt file using the infile statement in a data step**

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\* Importing the Bone.txt data from the P: directory to the H: directory.\*/

/\* This data set contains only numerical data and is comma delimited. \*/

/\* This is the first Bone data set used in this exercise. \*/

/\* The infile statement in the data step is used to import the data. \*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

**data** morne.Bone\_Txt;

infile "P:\ebw\2019\Statistiek\Nagraadse Modules\MVAR(10600721)\Data\BONE.txt" dlm=',' firstobs=**2**;

input Obs x1 x2 x3 x4;

**run**;

**proc** **print** data=morne.Bone\_Txt;

var Obs x1 x2 x3 x4;

**run**;

**Scenario 2:**

**Creating multiple data sets using the output statement**

**in a data step**

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\* The Bone\_Txt data set is now a SAS file stored in the morne library. \*/

/\* Once a data set is a SAS file we can use the set statement to call the\*/

/\* data. The Bone\_Txt is split into three data sets in as follows. \*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

**data** morne.Bone1 morne.Bone2 morne.Bone3;

set morne.Bone\_Txt ;

if Obs < **10** then output morne.Bone1;

else if **10**<=Obs<**15** then output morne.Bone2;

else if Obs>=**15** then output morne.Bone3;

**run**;

**proc** **print** data=morne.Bone1;

var Obs x1 x2 x3 x4;

**run**;

**proc** **print** data=morne.Bone2;

var Obs x1 x2 x3 x4;

**run**;

**proc** **print** data=morne.Bone3;

var Obs x1 x2 x3 x4;

**run**;

**Scenario 3:**

**Importing a prn file using the infile statement in a data step**

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\* Importing my BoneLabels.prn data from the P: directory to the H: \*/

/\* directory. This data set contains characters and is space delimited. \*/

/\* This is the second Bone Data set used in this exercise. \*/

/\* The infile statement in the data step is used to import the data. \*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

**data** morne.Bone\_Prn;

infile "P:\ebw\2019\Statistiek\Nagraadse Modules\MVAR(10600721)\Data\BoneLabels.prn" dlm=' ' firstobs=**2**;

input Obs Province $ Language $;

**run**;

**proc** **print** data=morne.Bone\_Prn;

var Obs Province Language;

**run**;

**Scenario 4:**

**Two data sets are merged and used to create further data sets**

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\* Merging the two Bone data sets (Bone\_prn and Bone\_txt) using the \*/

/\* merge statement. Both these files are now SAS files. \*/

/\* Note that the two data sets have column OBS in common. \*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

**data** morne.Bone\_Merge;

merge morne.Bone\_txt morne.Bone\_prn;

by Obs;

**run**;

**proc** **print** data=morne.Bone\_Merge;

**run**;

/\* Splitting the merged Bone data set (Bone\_Merge) into the different \*/ /\* Provinces using the output statement. \*/

/\* View the four data sets in your explorer. \*/

**data** morne.AFR morne.ENG morne.German morne.Italian;

set morne.Bone\_Merge;

if Language = 'AFR' then output morne.AFR;

else if Language = 'ENG' then output morne.ENG;

else if Language = 'GERMAN' then output morne.German;

else if Language = 'ITALIAN' then output morne.Italian;

**run**;

/\* Keeping only the observations from province=NC in Bone\_Merge dataset. \*/

/\* A new dataset NC\_Province is created using the where statement. \*/

/\* The where statement can only we used on a SAS dataset. \*/

/\* View the dataset in your explorer. \*/

**data** morne.NC\_Province;

set morne.Bone\_Merge;

where Province = 'NC';

**run**;

**Scenario 5:**

**Importing data set from the default directory using the**

**path statement in the data step**

**data** morne.region;

infile "&path\REGION.txt" dlm="&" firstobs=**2**;

input Region$ State$ Month$ Expense Revenue;

format Expense Dollar10.2 Revenue Dollar10.

run;

**proc** **print** data=morne.region;

var Region State Month Expense Revenue;

**run**;

**Scenario 6:**

**Importing a txt data set using proc import**

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\* Importing the Region data from the P: directory. \*/

/\* Further manipulations can be done using the data step. \*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

title 'The region dataset';

options nodate ps=**60** ls=**80**;

**proc** **import** datafile="P:\ebw\2019\Statistiek\Nagraadse Modules\MVAR(10600721)\Data\REGION.txt" out=morne.region dbms=dlm replace;

delimiter='&';

getnames=yes;

**run**;

**proc** **print** data=morne.region;

**run**;

**data** morne.region2;

set morne.region;

Profit=Revenue-Expenses;

format Month mmddyy10. Profit Dollar10.;

**run**;

**proc** **print** data=morne.region2;

sum Expenses Revenue Profit;

**run**;

**Scenario 7:**

**Exporting a SAS data set as html, pdf or rtf files**

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\* The Bone\_Merge file is being exported here. Some manipulations are \*/

/\* first done before the data is exported \*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

ods listing close;

ods html file="H:\Honours\SAS\_tutorial\myreport.html";

ods pdf file="H:\Honours\SAS\_tutorial\myreport.pdf";

ods rtf file=" H:\Honours\SAS\_tutorial\myreport.rtf";

/\* or your could use the path statement \*/

\* ods html file="&path\myreport.html";

\* ods pdf file="&path\myreport.pdf";

\* ods rtf file="&path\myreport.rtf";

**proc** **freq** data=morne.Bone\_Merge;

title 'Report 1';

tables Province Language Province\*Language;

**run**;

**proc** **means** data=morne.Bone\_Merge;

title 'Report 2';

var x1 x2 x3 x4;

**run**;

**proc** **print** data= morne.Bone\_Merge;

title 'Report 3';

var Obs x1 x2 x3 x4 Language Province;

where x4 <= **50**;

sum x1 x2 x3 x4;

**run**;

ods \_all\_ close;

ods listing;

title;

**Exporting a SAS data set as a csv file**

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\* The SAS file Bone\_Merge file is being exported here as a csv file. \*/

/\* The BoneMerge.csv can be opened in Excel. \*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

**proc** **export** data=morne.Bone\_merge

outfile="H:\Honours\SAS\_tutorial\BoneMerge.csv"

dbms=csv

replace;

**run**;

\*\*\*\*\*\*\* END OF LESSON 1 \*\*\*\*\*\*\*

**Lesson 2: Univariate Statistical Analyses**

**Proc steps for basic statistical analyses**

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\* The Irisdata.txt is imported for analysis. This data set contains \*/

/\* 150 observations and 4 variables. Three Iris species are used, thus \*/

/\* 50 observations per specie. \*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

**data** morne.Iris\_data;

infile "P:\ebw\2019\Statistiek\Nagraadse Modules\MVAR(10600721)\Data\IrisData.txt" dlm=" " ;

input Sepal\_Length Sepal\_Width Petal\_Length Petal\_Width Group;

if Group=**2** then Specie="Iris versicolor";

if Group=**3** then Specie="Iris virginica";

else if Group=**1** then Specie="Iris setosa";

drop Group;

**run**;

**The proc sort step for sorting data**

/\* It is important to note that the Iris\_data needs to be sorted \*/

/\* according to specie for some of the next procedures. \*/

/\* The proc sort step can be used to sort this data \*/

**proc** **sort** data=morne.Iris\_data out=morne.Sorted\_Iris;

by Specie;

**run**;

**proc** **print** data=morne.Sorted\_Iris;

**run**;

**The proc sgplot step for graphical display of the variables**

/\* Creating a Scatterplot for two variables. \*/

**proc** **sgplot** data=morne.Iris\_data;

title 'Scatter plot of the Iris data';

scatter y=Sepal\_Length x=Sepal\_Width / group=Specie;

**run**;

/\* Creating a Boxplot for a variable. \*/

**proc** **sgplot** data=morne.Iris\_data;

title 'Boxplot of the Iris data';

hbox Sepal\_Length / category=Specie;

**run**;

/\* Creating a Scatterplot and Ellipse for two variables by Specie. \*/

**proc** **sgplot** data=morne.Sorted\_Iris;

title 'Scatter plot of the Iris data by Specie';

by Specie;

scatter y=Sepal\_Length x=Sepal\_Width / group=Specie;

ellipse y=Sepal\_Length x=Sepal\_Width / alpha=**0.1**;

**run**;

/\* Creating a Histogram for a variables by Specie. \*/

**proc** **sgplot** data=morne.Sorted\_Iris;

title 'Histogram of the Iris data by Specie';

by Specie;

histogram Sepal\_Length;

density Sepal\_Length;

density Sepal\_Length / type=kernel;

keylegend / location=outside position=topmiddle;

**run**;

**The proc means step to obtain summary statistics**

/\* Obtaining the summary statistics can done using the proc means step \*/

**proc** **means** data=morne.Sorted\_Iris n mean max min range var std sum LCLM UCLM SKEW KURT fw=**8**;

title 'Means of the Iris data by Specie';

by Specie;

var Sepal\_Length Sepal\_Width Petal\_Length Petal\_Width;

output out=morne.Summary\_Stats;

**run**;

**proc** **print** data=morne.Summary\_Stats;

title "Summary Statistics of the Iris Data";

**run**;

**The proc univariate step to perform univariate normality tests**

**and confidence intervals**

/\* Performing a normality test on the Iris data set for each specie \*/

title 'Univariate normal test of the Iris data by Specie';

ods graphics off;

**proc** **univariate** data=morne.Sorted\_Iris normaltest;

by Specie;

var Sepal\_Length Sepal\_Width Petal\_Length Petal\_Width;

histogram Sepal\_Length Sepal\_Width Petal\_Length Petal\_Width;

qqplot Sepal\_Length Sepal\_Width Petal\_Length Petal\_Width / normal(mu=est sigma=est color=red l=**2**)square;

**run**;

/\* Creating confidence intervals for parameters by specie \*/

title 'Confidence intervals of Iris data by Specie';

ods select BasicIntervals;

**proc** **univariate** data=morne.Sorted\_Iris cibasic(alpha=**.1**);

by Specie;

var Sepal\_Length Sepal\_Width Petal\_Length Petal\_Width;

**run**;

**The proc corr step to perform a correlation analysis,**

**obtaining confidence regions for bivariate data**

**and the mean vector**

/\* Creating four different correlations with the proc corr step \*/

title 'Correlations of the of Iris data by Specie';

ods graphics on;

**proc** **corr** data=morne.Sorted\_Iris pearson spearman kendall hoeffding plots=matrix(histogram);

by Specie;

var Sepal\_Length Sepal\_Width;

\* with Petal\_Length Petal\_Width;

**run**;

ods graphics off;

/\* Creating a 95% and 90% confidence region for the data using proc corr \*/

title 'Confidence region for the data';

ods graphics on;

**proc** **corr** data=morne.Sorted\_Iris plots=scatter(nvar=**4** alpha=**.05** **.1**);

by Specie;

var Sepal\_Length Sepal\_Width Petal\_Length Petal\_Width;

**run**;

ods graphics off;

/\* Creating a 95% and 90% confidence region for the mean using proc corr \*/

title 'Confidence region for the mean';

ods graphics on;

**proc** **corr** data=morne.Sorted\_Iris plots=scatter(ellipse=confidence nvar=**4** alpha=**.05** **.1**);

by Specie;

var Sepal\_Length Sepal\_Width Petal\_Length Petal\_Width;

**run**;

ods graphics off;

**The proc ttest step to perform t-test for a single mean**

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\* Doing an hypothesis test for the mean using proc ttest. \*/

/\* Since we are doing it by Specie, we need to use the sorted data set \*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

title 'Hypothesis test: T-test';

ods graphics on;

**proc** **ttest** data=morne.Sorted\_Iris alpha=**0.05** H0=**5**;

by Specie;

var Sepal\_Length;

**run**;

ods graphics off;

\*\*\*\*\*\*\* END OF LESSON 2 \*\*\*\*\*\*\*

**Lesson 3: Multivariate Statistical Analyses**

**Proc steps for ANOVA and generalized linear models**

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\* The Irisdata.txt is imported for analysis. This data set contains \*/

/\* 150 observations and 4 variables. Three Iris species are used, thus \*/

/\* 50 observations per specie. \*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

**data** morne.Iris\_data;

infile "P:\ebw\2019\Statistiek\Nagraadse Modules\MVAR(10600721)\Data\IrisData.txt" dlm=" " ;

input Sepal\_Length Sepal\_Width Petal\_Length Petal\_Width Group;

if Group=**2** then Specie="Iris versicolor";

if Group=**3** then Specie="Iris virginica";

else if Group=**1** then Specie="Iris setosa";

drop Group;

**run**;

/\* Note that for this exercise the data does not have to be sorted ! \*/

**The proc anova step to perform a univariate and multivariate analysis of variance**

title 'MANOVA, ANOVA and Tukeys test - Iris data';

ods graphics off;

**proc** **anova** data=morne.Iris\_data;

class Specie;

model Sepal\_Length Sepal\_Width Petal\_Length Petal\_Width = Specie;

manova h=Specie / printh printe;

means Specie / tukey;

**run**;

ods graphics on;

**The proc glm step to perform a univariate and multivariate**

**analysis of variance**

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\* The proc glm step is used to perform the ANOVA, MANOVA and T-tests \*/

/\* The output of the analysis is written to another data set. \*/

/\* This output can now be used in further analyses like normality tests \*/

/\* and graphical displays \*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

title 'MANOVA, ANOVA and Tukey, LSD and Boneferroni tests - Iris data';

ods graphics on;

**proc** **glm** data=morne.Iris\_data alpha=**0.1**;

class Specie;

model Sepal\_Length Sepal\_Width Petal\_Length Petal\_Width = Specie / ss3;

manova h=Specie / printe printh;

means Specie / LSD Tukey Bon;

output out=morne.ANOVA COOKSD=Distance RSTUDENT=Residuals PRESS=Press H=Leverage DFFITS=DFfits;

**run**;

ods graphics off;

**proc** **print** data=morne.ANOVA;

var Distance Residuals Press Leverage DFfits;

**run**;

/\* Normality test of the residuals. \*/

title 'Univariate normal test of the Residuals';

**proc** **univariate** data=morne.ANOVA normaltest;

var Residuals;

histogram Residuals;

qqplot Residuals / normal(mu=est sigma=est color=red l=**2**)square;

**run**;

**Proc steps for regression analyses**

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\* Preparing a data set for regression analysis. \*/

/\* A dummy variable is created, Transformations and Summations are \*/

/\* performed on the original variables. \*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

**data** morne.Regress;

input Y1 Y2 X1 X2 X3 X4 X5 @@;

if Y1<**50** then dummy=**1**;

else dummy=**0**;

X1\_square=X1\*\***2**;

logY2=log(Y2);

SumX1\_X5=sum(X1,X2,X3,X4,X5);

datalines;

44 89.47 44.609 11.37 62 178 182 40 75.07 45.313 10.07 62 185 185

44 85.84 54.297 8.65 45 156 168 42 68.15 59.571 8.17 40 166 172

38 89.02 49.874 9.22 55 178 180 47 77.45 44.811 11.63 58 176 176

51 67.25 45.118 11.08 48 172 172 54 91.63 39.203 12.88 44 168 172

51 73.71 45.790 10.47 59 186 188 57 59.08 50.545 9.93 49 148 155

49 76.32 48.673 9.40 56 186 188 48 61.24 47.920 11.50 52 170 176

52 82.78 47.467 10.50 53 170 172

;

**proc** **print** data=morne.Regress;

**run**;

**The proc reg step for multiple linear regression analysis**

title 'Confidence band and diagnostic plots';

ods graphics on;

**proc** **reg** data=morne.Regress corr simple alpha=**0.01**

plots=diagnostics(stats=(default aic sbc cp bic))

plots(only label)=(RStudentByLeverage CooksD);

model Y1 = X1 X2 X3 X4 X5 / clm cli;

plot rstudent.\*p.;

plot nqq.\* rstudent.;

output out=morne.Results student=resid;

**run**;

ods graphics off;

/\* Testing for normality in residuals. \*/

**proc** **univariate** data=morne.Results normaltest;

title 'Univariate normal test on the residuals';

var Residuals;

**run**;

/\* Peforming Variable selection, Mallows Cp, AIC, BIC. \*/

ods graphics on;

**proc** **reg** data=morne.Regress;

title 'Variable selection using different criteria';

model Y1 = X1 X2 X3 X4 X5 / selection=rsquare cp aic bic;

plot cp.\*np. / cmallows=blue;

**run**;

ods graphics off;

/\* Checking for influential cases. \*/

**proc** **reg** data=morne.Regress;

title 'Influential cases';

model Y1 = X1 X2 X3 X4 X5 / influence;

**run**;

/\* Peforming a Stepwise regression. \*/

**proc** **reg** data=morne.Regress;

title 'Stepwise regression';

model Y1 = X1 X2 X3 X4 X5 / selection=forward;

model Y1 = X1 X2 X3 X4 X5 / selection=backward;

**run**;

/\* Testing for multicollinearity. \*/

**proc** **reg** data=morne.Regress;

title 'Multicollinearity';

model Y1 = X1 X2 X3 X4 X5 / collin tol vif;

**run**;

**The proc glm step for**

**multivariate multiple linear regression analysis**

/\* Peforming a Multivariate multiple regression. \*/

**proc** **glm** data=morne.Regress alpha=**0.1**;

title 'Multivariate regression';

model Y1 Y2 = X1 X2 X3 X4 X5 / ss3;

manova h= X1 X2 X3 X4 X5 / printe printh;

**run**;

**Proc steps for discriminant and classification analysis**

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\* Two data sets, a learning and test data set, will be used here. \*/

/\* The Hemophilia data is the data set of concern. A discriminant model \*/

/\* is built on the learning data and this model is used to classify the \*/

/\* test data. \*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\* Importing the learning sample (HemoPhiliaLearn.txt) to build model. \*/

title 'The Hemophilia data';

**data** morne.HemoLearn;

infile 'P:\ebw\2019\Statistiek\Nagraadse Modules\MVAR(10600721)\Data\HemophiliaLearn.txt' delimiter=' ';

input Class AHF\_activity AHF\_antigen;

if class=**2** then Group='Obligatory\_Carriers';

else if class=**1** then Group='Noncarriers';

drop Class;

**run**;

**proc** **print** data=morne.HemoLearn;

**run**;

/\* Importing the test sample (HemophiliaTest.txt) for classification. \*/

**data** morne.HemoTest;

infile 'P:\ebw\2019\Statistiek\Nagraadse Modules\MVAR(10600721)\Data\HemophiliaTest.txt' delimiter=' ';

input AHF\_activity AHF\_antigen;

**run**;

**proc** **print** data=morne.HemoTest;

**run**;

**The proc discrim step for discriminant and classification analysis**

/\* Performing a Linear discriminant analysis (LDA) (pool=yes). \*/

**proc** **discrim** data=morne.HemoLearn outstat=morne.LinearModel method=normal pool=yes crossvalidate list;

title2 'Linear discriminant analysis';

class Group;

var AHF\_activity AHF\_antigen;

**run**;

/\* Classifying the test data with Linear Model. \*/

**proc** **discrim** data=morne.LinearModel testdata=morne.HemoTest testout=morne.LinearTestOut testlist;

title2 'Classifying the test data with Linear model';

class Group;

var AHF\_activity AHF\_antigen;

**run**;

**proc** **print** data=morne.LinearTestOut;

**run**;

/\* Exproting the LDA output to a csv file. \*/

**proc** **export** data=morne.LinearTestOut

outfile="H:\Honours\SAS\_tutorial\LDAOutput.csv"

dbms=csv

replace;

**run**;

/\* Performing a Quadratic discriminant analysis (QDA) (pool=no). \*/

**proc** **discrim** data=morne.HemoLearn outstat=morne.QuadraticModel method=normal pool=no crossvalidate list;

title2 'Quadratic discriminant analysis';

class Group;

var AHF\_activity AHF\_antigen;

**run**;

/\* Classifying the test data with Quadratic Model \*/

**proc** **discrim** data=morne.QuadraticModel testdata=morne.HemoTest testout=morne.QuadraticTestOut testlist;

title2 'Classifying the test data with Quadratic model';

class Group;

var AHF\_activity AHF\_antigen;

**run**;

**proc** **print** data=morne.QuadraticTestOut;

**run**;

/\* Performing a Non-parametric (Kernel Density) discriminant analysis. \*/

**proc** **discrim** data=morne.HemoLearn outstat=morne.KernelModel method=npar metric=full r=**4** crossvalidate list;

title2 'Kernel density discriminant analysis';

class Group;

var AHF\_activity AHF\_antigen;

**run**;

/\* Classifying the test data with Non-parametric Model. \*/

**proc** **discrim** data=morne.KernelModel testdata=morne.HemoTest testout=morne.KernelTestOut testlist;

title2 'Classifying the test data with Kernel density Non-parametric model';

class Group;

var AHF\_activity AHF\_antigen;

**run**;

**proc** **print** data=morne.KernelTestOut;

**run**;

/\* Performing Non-parametric (K-Nearest Neighbour) discriminant analysis.\*/

**proc** **discrim** data=morne.HemoLearn outstat=morne.KNNModel method=npar metric=full k=**6** crossvalidate list;

title2 'KNN discriminant analysis';

class Group;

var AHF\_activity AHF\_antigen;

**run**;

/\* Classifying the test data with Non-parametric Model \*/

**proc** **discrim** data=morne.KNNModel testdata=morne.HemoTest testout=morne.KNNTestOut testlist;

title2 'Classifying the test data with KNN Non-parametric model';

class Group;

var AHF\_activity AHF\_antigen;

**run**;

**proc** **print** data=morne.KNNTestOut;

**run**;

**The proc logistic step for logistic regression analysis**

/\* Performing a logistic regression on the Hemophilia data set. \*/

**proc** **logistic** data=morne.HemoLearn desc outest=morne.Betas covout alpha=**0.01**;

title 'Logistic regression analysis';

model Group=AHF\_activity AHF\_antigen / expb;

output out=morne.LogisticPred p=phat lower=lcl upper=ucl

predprob=(individual crossvalidate);

**run**;

**proc** **print** data=morne.Betas;

title2 'Parameter Estimates and Covariance Matrix';

**run**;

**proc** **print** data=morne.LogisticPred;

title2 'Predicted Probabilities and 99% Confidence Limits';

**run**;

/\* Exporting the output of logistic regression to a csv file \*/

**proc** **export** data=morne.LogisticPred

outfile="H:\Honours\SAS\_tutorial\LogisticOutput.csv"

dbms=csv

replace;

**run**;

**The proc stepdisc step for variable selection in discriminant analysis**

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\* The Irisdata.txt is imported to for analysis. This data set contains \*/

/\* 150 observations and 4 variables. Three Iris species are used, thus \*/

/\* 50 observations per specie. \*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

**data** morne.Iris\_data;

infile "P:\ebw\2019\Statistiek\Nagraadse Modules\MVAR(10600721)\Data\IrisData.txt" dlm=" " ;

input Sepal\_Length Sepal\_Width Petal\_Length Petal\_Width Group;

if Group=**2** then Specie="Iris versicolor";

if Group=**3** then Specie="Iris virginica";

else if Group=**1** then Specie="Iris setosa";

drop Group;

**run**;

/\* Variable selection in discriminant analysis using the Iris data. \*/

**proc** **stepdisc** data=morne.Iris\_data bsscp tsscp;

title 'Variable selection in discriminant analysis';

class Specie;

var Sepal\_Length Sepal\_Width Petal\_Length Petal\_Width;

**run**;

**Proc steps for ordination analysis**

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\* Importing the Bone.txt data from the P: directory to the H: directory.\*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

**data** morne.Bone\_Txt;

infile "P:\ebw\2019\Statistiek\Nagraadse Modules\MVAR(10600721)\Data\BONE.txt" dlm=',' firstobs=**2**;

input Obs x1 x2 x3 x4;

**run**;

**proc** **print** data=morne.Bone\_Txt;

var Obs x1 x2 x3 x4;

**run**;

**The proc princomp step for principal component analysis**

/\* PCA on the Bone data using the covariance matrix. \*/

**proc** **princomp** data=morne.Bone\_txt cov out=morne.PCA\_cov\_Bone;

title 'Principal component analysis (PCA) using the covariance matrix';

var x1 x2 x3 x4;

**run**;

**proc** **print** data=morne.PCA\_cov\_Bone;

title 'Data and PCA scores';

**run**;

title 'Plot of PCA scores: Covariance';

%***plotit***(data=morne.PCA\_cov\_Bone, plotvars=Prin2 Prin1, color=black, colors=blue);

**run**;

/\* Exporting the output of PCA to a csv file. \*/

**proc** **export** data=morne.PCA\_cov\_Bone

outfile="H:\Honours\SAS\_tutorial\PCA\_cov\_Output.csv"

dbms=csv

replace;

**run**;

/\* PCA on the Bone data using the correlation matrix. \*/

**proc** **princomp** data=morne.Bone\_txt out=morne.PCA\_cor\_Bone;

title 'Principal component analysis (PCA) using the correlation matrix';

var x1 x2 x3 x4;

**run**;

**proc** **print** data=morne.PCA\_cor\_Bone;

title 'Data and PCA scores';

**run**;

title 'Plot of PCA scores: Correlation';

%***plotit***(data=morne.PCA\_cor\_Bone, plotvars=Prin2 Prin1, color=black, colors=blue);

**run**;

/\* Exporting the output of PCA to a csv file \*/

**proc** **export** data=morne.PCA\_cor\_Bone

outfile="H:\Honours\SAS\_tutorial\PCA\_cor\_Output.csv"

dbms=csv

replace;

**run**;

**The proc factor step for factor analysis**

/\* Principal component method using the covariance matrix. \*/

**proc** **factor** res data=morne.Bone\_txt method=prin out=morne.Fact\_Cov\_Prin cov nfact=**2** rotate=varimax score scree preplot plot;

title 'Principal component method:Covariance matrix';

var x1 x2 x3 x4;

**run**;

/\* Principal component method using the correlation matrix. \*/

**proc** **factor** res data=morne.Bone\_txt method=prin corr nfact=**2** rotate=varimax score out=out=morne.Fact\_Cor\_Prin;

title 'Principal component method:Correlation matrix';

var x1 x2 x3 x4;

**run**;

**proc** **print** data=morne.Fact\_Cov\_Prin;

title 'Data and Factor scores: Covariance matrix';

**run**;

**proc** **print** data=morne.Fact\_Cor\_Prin;

title 'Data and Factor scores: Correlation matrix';

**run**;

/\* Maximum likelihood method. \*/

**proc** **factor** data=morne.Bone\_txt method=ml hey corr nfact=**2** rotate=varimax score out=morne.Fact\_MaxL ;

title 'Maximum likelihood:Correlation matrix';

var x1 x2 x3 x4;

**run**;

**proc** **print** data=morne.Fact\_MaxL;

title 'Data and Factor scores: Maximum Likelihood';

**run**;

/\* Exporting the output of Factor analysis output (Maximum likelihood) \*/

**proc** **export** data=morne.Fact\_MaxL

outfile="H:\Honours\SAS\_tutorial\FactMaxl.csv"

dbms=csv

replace;

**run**;

**The proc cancorr step for cananonical correlation analysis**

/\* Canonical correlation analysis (CCA)on the Bone data. \*/

**proc** **cancorr** data=morne.Bone\_txt all corr out=morne.CanCorrOutput vprefix=U\_vars wprefix=V\_vars;

title 'CCA:Correlation matrix';

var x1 x2 ;

with x3 x4 ;

**run**;

**proc** **print** data=morne.CanCorrOutput;

**run**;

/\* Exporting the output of Canonical correlation analysis to a csv file. \*/

**proc** **export** data=morne.CanCorrOutput

outfile="H:\Honours\SAS\_tutorial\CanCorrOutput.csv"

dbms=csv

replace;

**run**;

**The proc cluster step for cluster analysis**

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\* The data is read in using datalines in the data step. \*/

/\* Hierarchical and model-based clustering is performed. \*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

**data** morne.Cluster;

input x1 x2 x3 x4 y$;

datalines;

5.1 3.5 1.4 0.2 a

4.9 3.0 1.4 0.2 b

4.7 3.2 1.3 0.2 c

4.6 3.1 1.5 0.2 d

5.0 3.6 1.4 0.2 e

5.4 3.9 1.7 0.4 f

4.6 3.4 1.4 0.3 g

5.0 3.4 1.5 0.2 h

4.4 2.9 1.4 0.2 i

4.9 3.1 1.5 0.1 j

5.4 3.7 1.5 0.2 k

4.8 3.4 1.6 0.2 l

4.8 3.0 1.4 0.1 m

4.3 3.0 1.1 0.1 n

5.8 4.0 1.2 0.2 o

5.7 4.4 1.5 0.4 p

5.4 3.9 1.3 0.4 q

;

/\* Hierarchical clustering methods \*/

/\* The Single linkage clustering \*/

ods graphics on;

**proc** **cluster** data=morne.Cluster method=single outtree=morne.SingleCluster;

title 'Hierarchical method: Single linkage';

var x1 x2 x3 x4;

id y;

**run**;

ods graphics off;

**proc** **tree** data=morne.SingleCluster;

**run**;

**proc** **print** data=morne.SingleCluster;

**run**;

/\* The Complete linkage clustering \*/

ods graphics on;

**proc** **cluster** data=morne.Cluster method=complete outtree=morne.ComplCluster;

title 'Hierarchical method: Complete linkage';

var x1 x2 x3 x4;

id y;

**run**;

ods graphics off;

/\* Model-based clustering methods \*/

ods graphics on;

**proc** **cluster** data=morne.Cluster method=eml outtree=morne.ModelCluster;

title 'Model based clustering method';

var x1 x2 x3 x4;

id y;

**run**;

ods graphics off;

\*\*\*\*\*\*\* END OF LESSON 3 \*\*\*\*\*\*\*