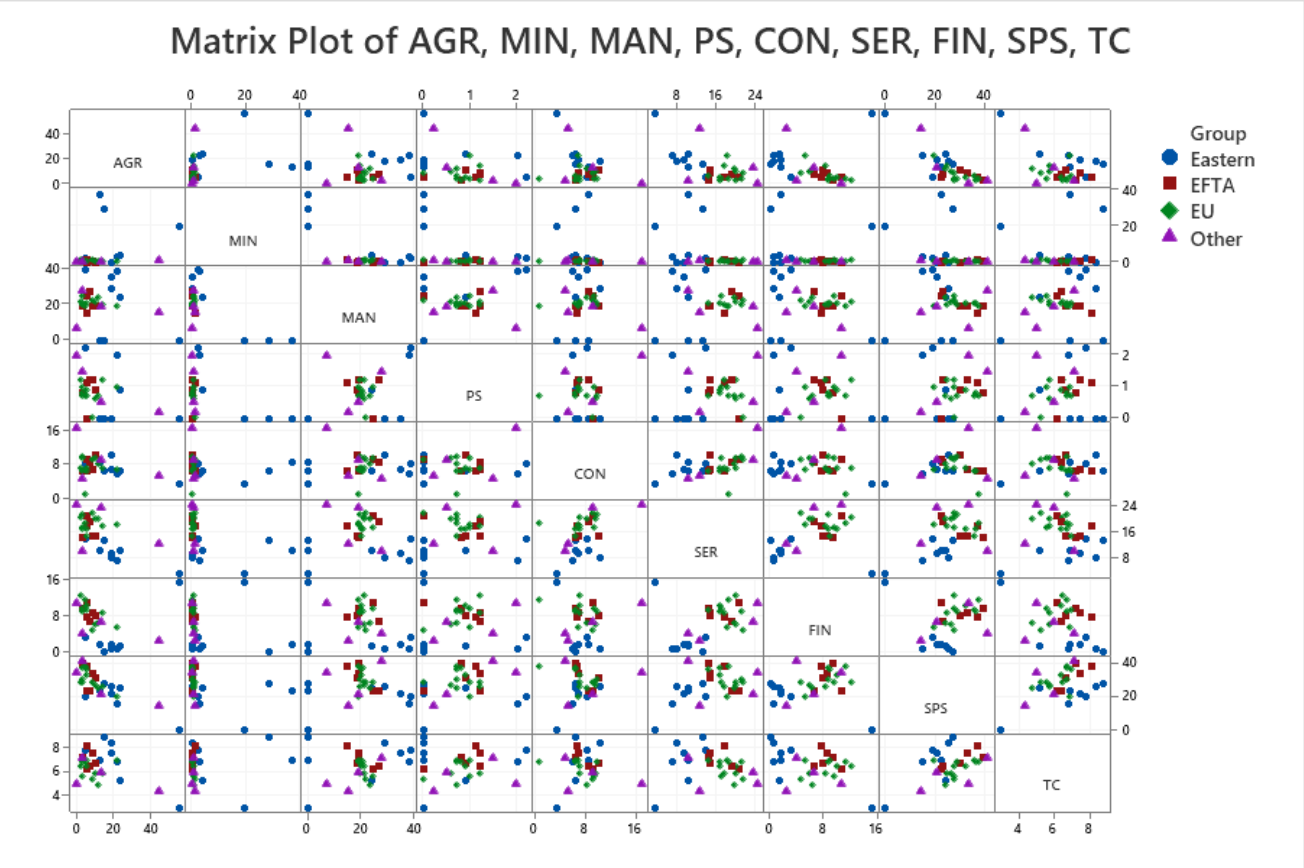
Additional Multivariate Exercise 2

Data: The European employment data were described in the Manly textbook. The data have similar structure as the Egyptian skulls data (Exercise 1), with some differences:

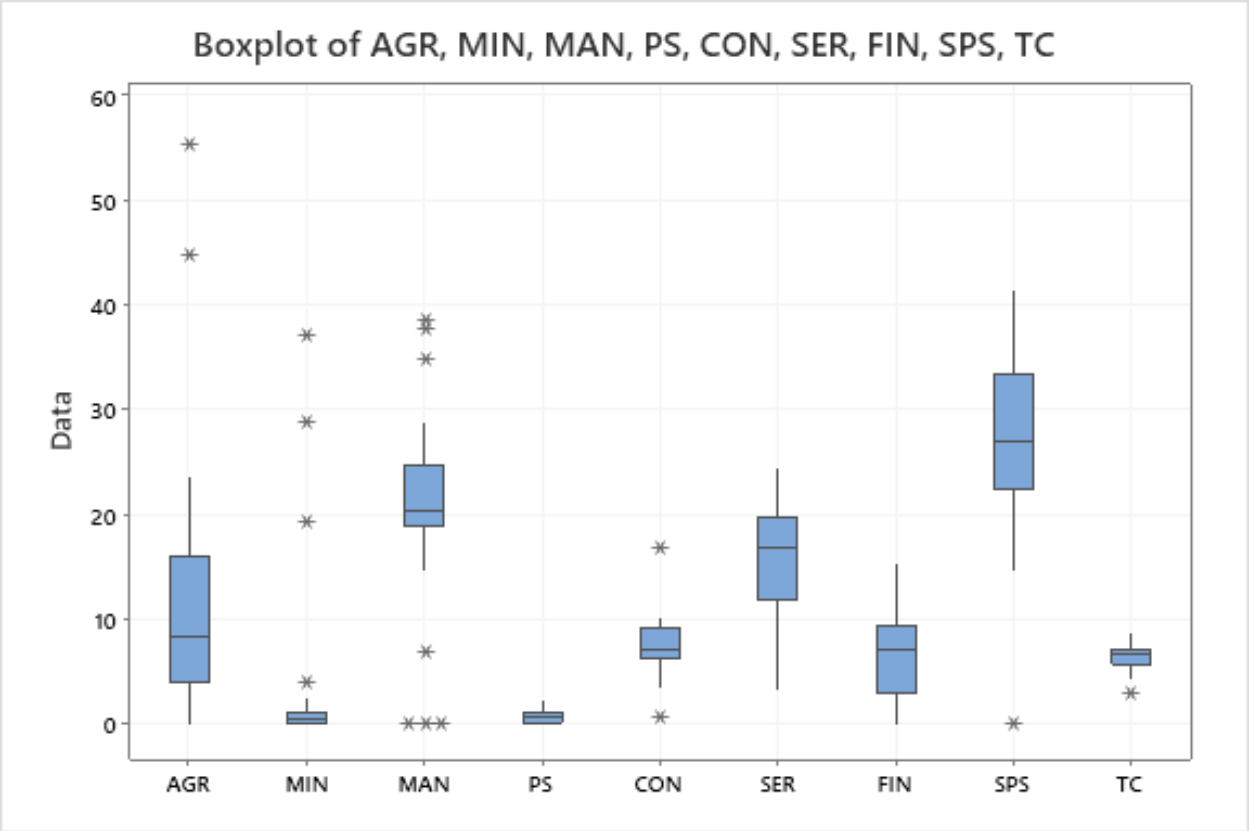
- the grouping (political group) is of less interest for these data, and the focus of analysis is more on considering the data for all 26 countries together than comparing the political groups,
- the variables are percentages out of a total rather than individual measurements.

First, by summing the values for the nine variables it is seen that the total indeed equals 100%, apart from some rounding error. This will in itself induce some negative correlations between variables; for example, if one variable is very large, the others cannot also be very large.

Initial graphical exploration: Also for these data a scatterplot matrix is the obvious starting point. Such a plot generated by Minitab is shown below.



The patterns across the pairwise plots are much more variable than for the Egyptian skulls data. It is also seen that some distributions appear far from symmetrical, and several extreme values within distributions are seen (e.g. for agriculture ("AGR") and mining ("MIN")). Some positive and negative associations between variables appear to exist, but many of the patterns are difficult to assess because of the irregular distributions of the values. Boxplots for each of the variables (across all countries) would be a natural next step, in order to better see skewness in distributions and outliers. Because the variables are on the same percentage scale, a plot with all variables is possible, although some distributions have a much more narrow range.



The boxplots demonstrate that summarizing the distributions by their mean and standard deviation will not be appropriate here. The listing below gives the standard descriptive statistics for each of the variables. Additionally, it may be informative to identify the most striking (suspected) outliers from the boxplot, simply by searching the dataset for the extreme values in the plot.

Statistics

Variable	N	Mean	StDev	Minimum	Q1	Median	Q3	Maximum	Skewness
AGR	30	12.19	12.31	0.00	3.98	8.45	16.10	55.50	2.20
MIN	30	3.45	8.87	0.00	0.10	0.50	1.10	37.30	3.11
MAN	30	20.29	9.46	0.00	18.93	20.30	24.72	38.70	-0.49
PS	30	0.800	0.621	0.000	0.150	0.800	1.200	2.200	0.46
CON	30	7.530	2.733	0.600	6.375	7.050	9.125	16.900	0.80
SER	30	15.637	5.160	3.300	11.875	16.800	19.875	24.500	-0.46
FIN	30	6.650	3.987	0.000	2.925	7.150	9.450	15.300	-0.05
SPS	30	26.99	8.73	0.00	22.47	27.00	33.40	41.60	-0.80
TC	30	6.453	1.233	3.000	5.750	6.750	7.200	8.800	-0.62

- agriculture high values: Albania, Turkey;
- mining high values: Czech/Slovakia, Hungary, Albania;
- manufacturing high values: Yogoslavia, Romania, Bulgaria, and the three countries with high values for mining have values of zero (suspect?);

- construction: high (Gibraltar) and low (Netherlands);
- services: zero value for Albania (suspect?).

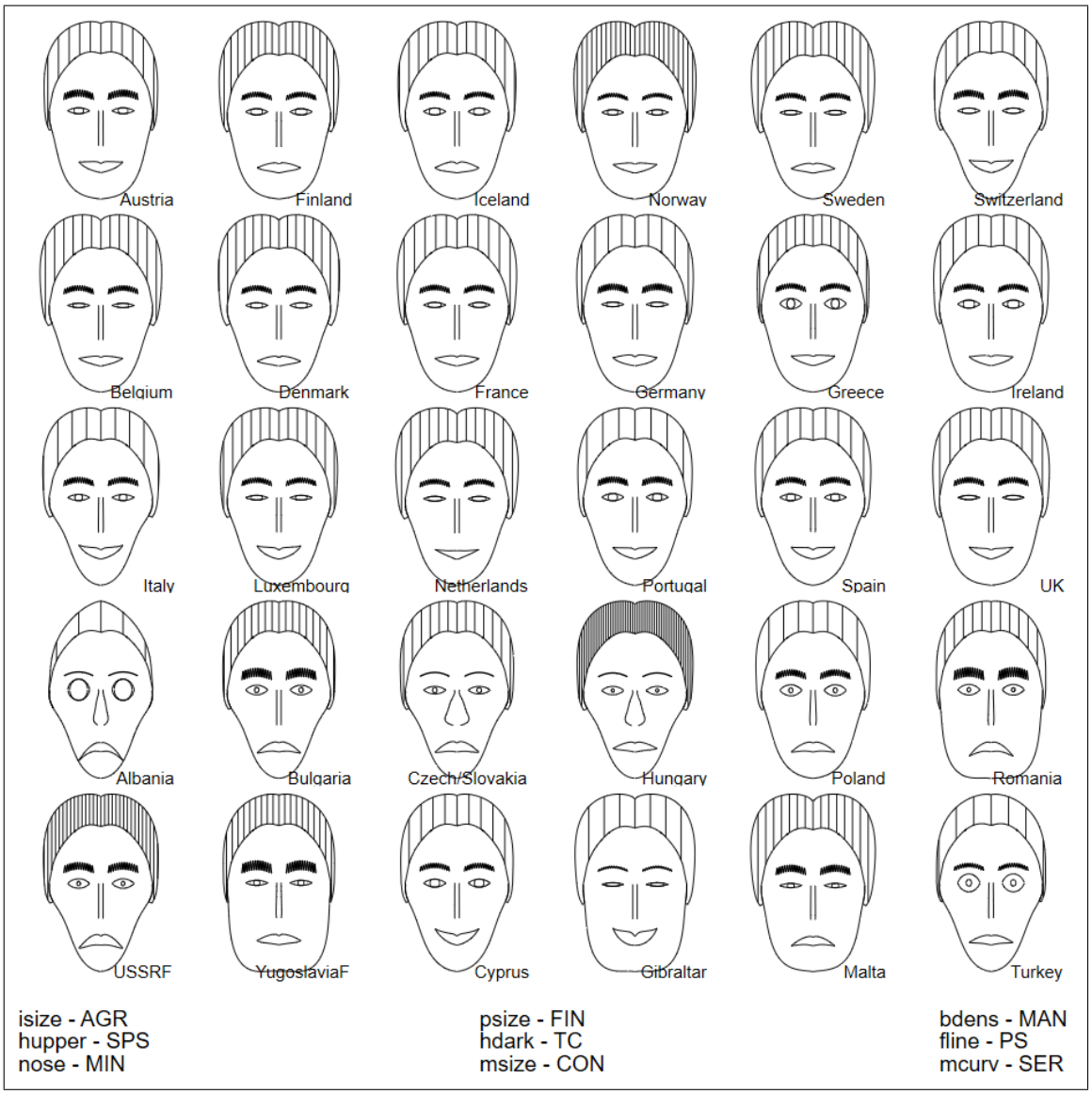
The list of extreme values identified three countries (Albania, Czech/Slovakia, Hungary) with unusual and possibly suspect values. As these will inevitably stand out in any attempt to create similar groups between groups, one may consider to remove them and focus on the other countries; for this solution, we will however keep them in.

The Pearson correlation coefficient will be substantially affected by extreme values and skewed distributions, therefore it is preferable to use the non-parametric Spearman rank correlation coefficient to quantify the association between variables. The listing below (from Stata) has manually been reduced from four decimals (the default) to two decimals, which should be sufficient for most practical purposes.

	agr	min	man	ps	con	ser	fin	sps	tc
agr	1.00								
min	0.32	1.00							
man	-0.11	-0.12	1.00						
ps	-0.40	0.06	0.24	1.00					
con	-0.19	-0.38	0.10	-0.05	1.00				
ser	-0.46	-0.32	-0.10	0.11	0.52	1.00			
fin	-0.58	-0.36	-0.27	0.15	0.05	0.46	1.00		
sps	-0.73	-0.36	-0.15	0.25	-0.03	0.24	0.42	1.00	
tc	-0.12	-0.12	0.09	0.07	-0.14	-0.36	-0.32	0.28	1.00

Most of the correlations substantially away from zero are negative, in particular with agriculture. There is however one positive correlation between construction and services.

Finally one version of a Chernoff plot will be shown. With nine variables there are many ways these can be mapped to facial features, resulting in visibly different plots. The plot below shows well both some of the outlying countries discussed above and the strong similarities between Western European (and in particular Scandinavian) countries.


Austria
Finland
Iceland
Norway
Sweden
Switzerland
Belgium
Denmark
France
Germany
Greece
Ireland
Italy
Luxembourg
Netherlands
Portugal
Spain
UK
Albania
Bulgaria
Czech/Slovakia
Hungary
Poland
Romania
USSRF
YugoslaviaF
Cyprus
Gibraltar
Malta
Turkey
isize - AGR
hupper - SPS
nose - MIN
psize - FIN
hdark - TC
msize - CON
bdens - MAN
fline - PS
mcurv - SER