Lab Assignment No. 4: Answer Key

1) What are the bivariate correlations for your data? What interpretations in the multivariate environment can you draw?

Since it is impossible to visualize your data beyond three dimensions, a series of bivariate analyses may sometimes prove informative when considering the multivariate interrelationships between variables. In this particular case a correlations matrix between your analyzed variables ought to be investigated in order to determine whether variables are related between one another and also if multicollinearity exists (indicated by extremely high correlations).

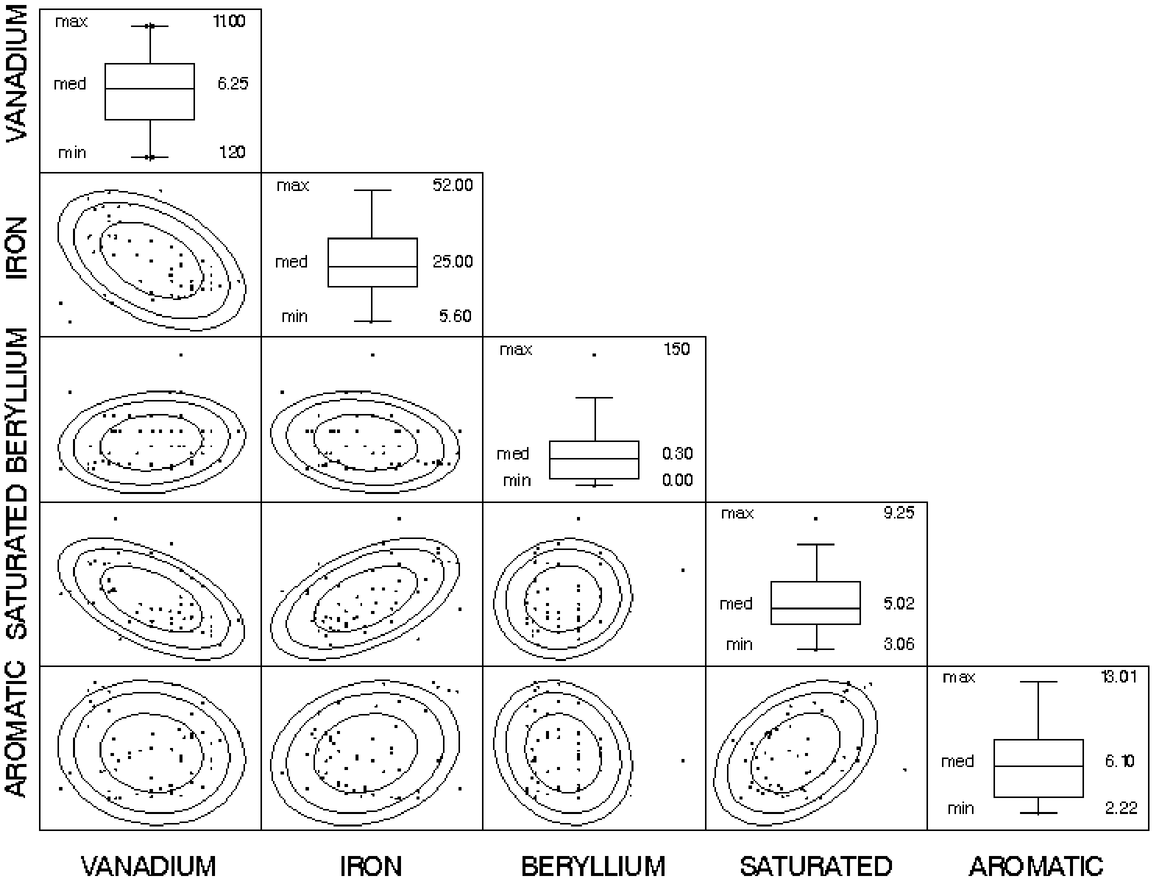
Table 1. Bivariate correlations for crude oil data (*p* values)

	Vanadium	Iron	Beryllium	Saturated	Aromatic
Vanadium	1				
Iron	-.44 *(< .001)*	1			
Beryllium	.11 *(.42)*	-0.16812 *(.22)*	1		
Saturated	-.55 *(< .001)*	.52 *(< .001)*	.07 *(.6)*	1	
Aromatic	-.08 *(.55)*	.18 *(.18)*	-.16 *(.23)*	.37 *(< .01)*	1

Also, investigating a scatterplot matrix of the relationships in Table 1 can provide additional information. As can be see in Figure 1 the correlations which were not significant produced almost completely circular confidence regions (confidence ellipsoids) around the data scatter. This substantiates that there is a weak relationship between the given variables (in this example *Saturated* and *Beryllium* for example).

Moreover, being able to investigate the 50%, 95% and 99% confidence ellipsoids for the bivariate relationships makes it possible to identify some bivariate (multivariate) outliers. As can be seen in Figure 1 there are at least several outliers in each of the plotted cells. However, since for the purposes of our analysis we would investigate not two pairings, but all variables considered at once, these outliers may not be outliers in the multivariate hyper space that we shall consider.

Figure 1. Scatterplot matrix with confidence ellipsoids



2) Provide the appropriate (most informative for further multivariate analyses) univariate measures of central tendency.

The appropriate univariate statistics here would be the means for each individual zone. Since our investigation is pertaining to comparing means between the three different zones at which crude oil is siphoned we want to inquire specifically into what the differences are in the chemical composition of the oil samples.

Table 2. Individual group means, standard deviations and standard errors

	Wilhelm			SubMuli			Upper		
Chemical	Mean	*S.E.*	*S.D.*	Mean	*S.E.*	*S.D.*	Mean	*S.E.*	*S.D.*
Vanadium	3.23	0.20	0.54	4.45	0.59	1.96	7.23	0.32	2.00
Iron	43.57	2.31	6.11	33.09	3.25	10.77	22.25	1.42	8.76
Beryllium	0.12	0.04	0.10	0.17	0.06	0.18	0.43	0.05	0.33
Saturated	6.80	0.33	0.87	6.56	0.38	1.27	4.66	0.16	0.99
Aromatic	11.54	0.53	1.41	5.48	1.04	3.44	5.77	0.38	2.37

Investigating the univariate statistics we can observe that iron is certainly the most prevalent chemical in crude oil. Also, crud oil siphoned at *Wilhelm* seems to contain most iron on average. It is also the crude oil with the most aromatic average carbon compounds (with the smallest deviation).

Vanadium on the other hand occurs most frequently in the *Upper* zones of siphoning. However, the amount of Vanadium content varies the most at this zone (followed closely by the crud oil from the *SubMuli* zone).

3) What is the conceptual difference between a Hotelling's T^2 and a Multivariate Analysis of Variance (MANOVA)?

Analogous to the univariate paired or independent samples *t*-tests Hotelling's T^2 compares the vector of means for two groups. Whereas in the univariate environment only one mean for each group is compared Hotelling's T^2 compares several means simultaneously between the two groups. The MANOVA is the generalization of the T^2 procedure where now not two but multiple group vectors of means can be compared. This is synonymous to the ANOVA *F*-test where the mean of a single variable is compared between three or more groups.

4) Assess whether your data is multivariate normally distributed and meets basic assumptions (Mardia's coefficients, Henze-Zirkler T, χ^2 Q-Q plot, and Bartlett's test).

Given that we are comparing across groups each individual group should be investigated if its data meet the multivariate normality assumption. One of the ways this has been demonstrated is to investigate the presence of multivariate outliers by looking at either leverage values or Mahalonobis distances. Since we are dealing with an unbalanced design, meaning different sample sizes within each one of the groups, the corresponding leverage critical values are going to be different for each zone. Table 3 summarized the three critical values discussed, the conservative value, the regression (including a constant) approach and the Raykov and Marcoulides proposed critical value based on the Mahalonobis distance χ^2 critical value, as well as the maximum observed leverage value within each group.

Table 3. Critical and maximum leverage values (by zone)

Zone	Conserv.	Constant	χ^2 Based	Max Leverage
SubMuli	0.91	1.09	1.60	0.89
Upper	0.26	0.32	0.43	0.43
Wilhelm	1.43	1.71	2.66	0.98

As can be seen from Table 3, given that we consider the χ^2 based critical value as most appropriate, none of the group's largest leverage values exceeded the critical value. If we were to consider the second critical value the highest leverage value in the *Upper* zone would be marginally significantly above the critical value. Given the difficult nature of multivariate normality we shall continue with the third critical value and thus not remove any observations.

As per multivariate normality Mardia's coefficients and the corresponding Henze-Zirkler tests were considered. Table 4 summarizes these estimates. Of the three zones only *Upper* appears to have a skew in its multivariate distribution. Accordingly the Henze-Zirkler test is significant, T = 6.25, p < .001. Again, given the difficult and complex nature of multivariate normality we shall only consider estimates with a probability of less than .01 as significant. Subsequently, the Henze-Zirkler test for the *SubMuli* zone was deemed non-significant, T = 2.25, p > .01.

Table 4. Multivariate normality coefficients

	Mardia's		Henze-Zirkler		Bartlett's	
Zone	Skewness	Kurtosis	*T*	*p*	χ^2	*p*
SubMuli	47.03	-0.47	2.25	0.02	49.13	0.02
	(.08)	*(.64)*				
Upper	92.62	1.88	6.25	< .001		
	(<.001)	*(.06)*				
Wilhelm	31.97	-1.49	0.76	0.45		
	(.62)	*(.14)*				

Further, the assumption of equal variance/covariance matrices was investigated using the Bartlett's test. From the χ^2 approximation, χ^2(30) = 49.13, p > .01, it can be inferred that the three siphoning zone's covariance matrices do not vary significantly from one another. Since the three groups do not have different variances/covariances and no multivariate outliers were detected the skew of the *Upper* group will be noted, but no further actions will be taken.

5) What hypotheses will your one-way MANOVA be testing?

In computing a MANOVA we are specifically investigating whether there are any differences between mean vectors of the analyzed variables between the specific independent groups. Therefore, we could express the hypotheses as follows:

$$H_0 : \underline{\bar{\mu}}_{SubMuli} = \underline{\bar{\mu}}_{Upper} = \underline{\bar{\mu}}_{Wilhelm}$$

$$H_A : Not; \underline{\bar{\mu}}_i \neq \underline{\bar{\mu}}_j$$

Basically what is being investigated is whether all of the group mean vectors are equivalent (the null hypothesis) or if there is a single variable pairing between groups that is not approximately equivalent.

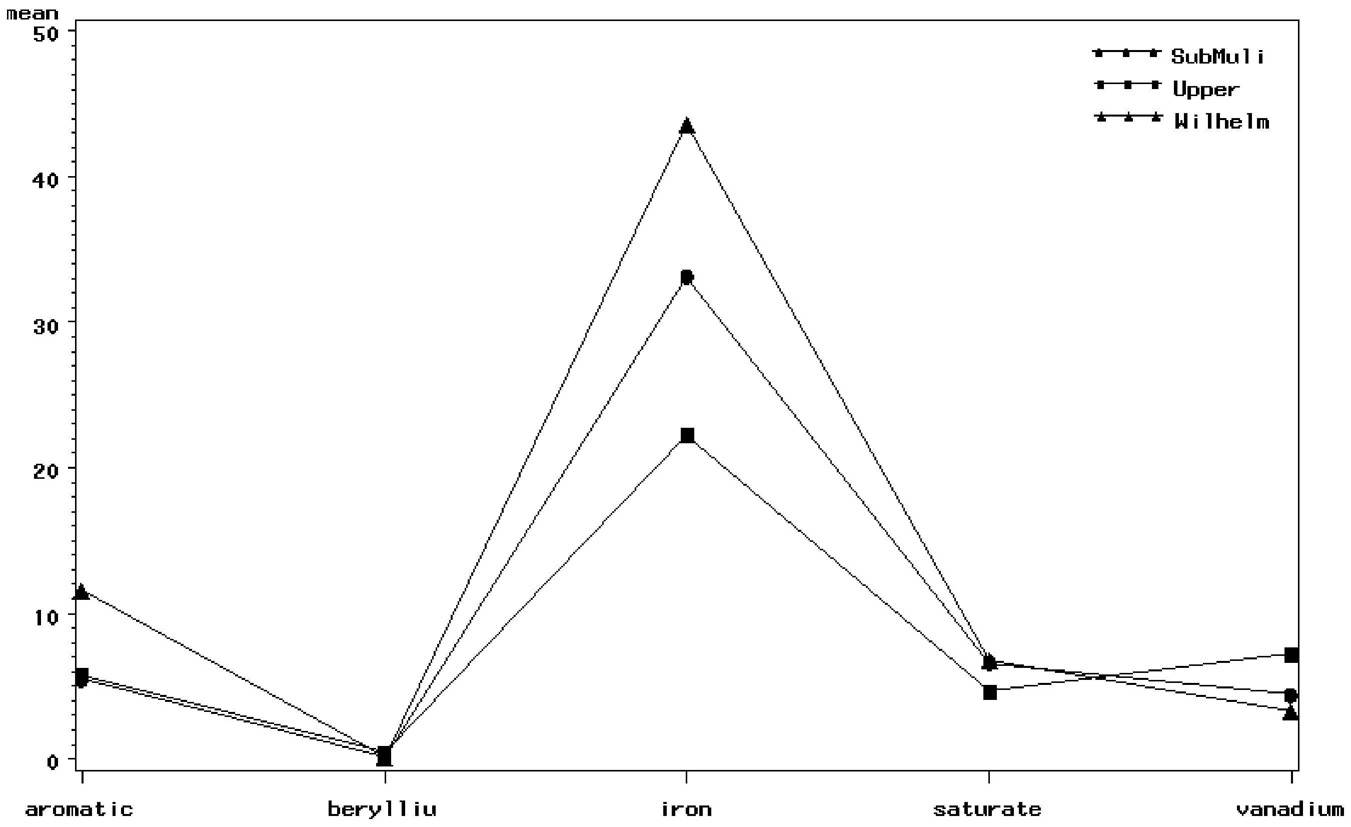
6) Run a one-way MANOVA. Report and discuss Wilk's Lambda and Roy's Maximum (or Largest) Root.

Using Wilk's lambda we were able to reject the null hypothesis that the mean vectors for the three groups are equivalent, Λ = .12, *F* (10, 98) = 18.98, *p* < .001. Further, we were able to arrive at the same conclusion using Roy's Largest Root as a decision criterion, $\lambda_{\max}$ = 4.18, *F* (5, 50) = 41.78, *p* < .001. This means that there exists at least one variable pairing between the groups of interest that do not have approximately equivalent means. However, in order to make specific inferences about the interrelationship between the chemicals and the different sites we need to follow-up the multivariate test wit ha series of univariate ANOVAs and *t*-tests.

7) What specific conclusions can you draw from the MANOVA results and the univariate ANOVA follow-up tests?

Investigating Figure 2 we can see that the means for the five chemicals vary across the three groups. It is also apparent that of the five chemicals iron is the most prevalent element in crude oil (given that it has the highest mean for all groups). The element Beryllium, on the other hand, seems to be problematic.

Figure 2. Means profile plot for chemical composition of crude oil by siphoning zones



It is difficult to infer differences from Figure 2 regarding Beryllium. Because the elements presence in crude oil is relatively small the differences on the provided scale cannot be teased apart from the profile plot. Instead, an ANOVA should be conduced for each element comparing the means of the element between the three siphoning groups.

Table 5 summarizes the five ANOVAs and gives their corresponding R-squared values. As can be seen there is a significant difference in at least one pairing of zones for each element (denoted by the fact that all ANOVAs produced significant *F* values).

Table 5. Univariate ANOVA follow-up tests

Chemical	df_1	df_2	F	p	R^2
Vanadium	2	53	19.17	<.001	.42
Iron	2	53	20.01	<.001	.43
Beryllium	2	53	5.88	<.01	.18
Saturated	2	53	22.67	<.001	.46
Aromatic	2	53	16.41	<.001	.46

What was earlier not possible (given only the multivariate test and the profile plot) can now be investigated using the ANOVA results. Though mean differences in Beryllium between the three groups seemed to be relatively small, at least one of the differences produced a significant *F* ratio. However, compared with the other four elements analyzed Beryllium does appear to be the one variable with the most ‘noise’ in the data. Again, it is not possible for us to make specific generalizations about the elements and the siphoning zones without further post-hoc investigations.

8) Using the simultaneous confidence intervals for the least square means what further conclusions can you make about the specific mean differences of the individual chemicals at each zone?

Even though the Roy-Bose simultaneous confidence intervals are often considered to be too conservative, they are appropriate when dealing with variable sets that are believed to have a linear relationship (i.e. where a linear combination between them exists). Since the five variables are chemical compositions of crude oil samples the linear-combination assumption is certainly present making the simultaneous confidence intervals a viable choice. Since we have found both Wilk’s Lambda and the follow-up ANOVAs to be significant we can now conduct post-hoc multiple comparisons. Even in the light of an omnibus test (here the one-way MANOVA) we still ought to guard ourselves again alpha inflation when evaluating the paired mean differences in our chemicals. Therefore the subsequent 12 *t*-test corresponding significance values are corrected (using a Bonferroni correction) to prevent Type I Error inflation.

Table 6. Univariate *t*-test multiple comparisons

			95% Simultaneous C.I.		
Chemical	Pair	Mean Difference	Lower	Upper	*p*
Vanadium	SubMuli-Upper	-2.78	-4.37	-1.19	< .001
	SubMuli-Wilhelm	1.22	-1.03	3.46	.56
	Upper-Wilhelm	4.00	2.08	5.91	< .001
Iron	SubMuli-Upper	10.84	3.28	18.39	< .01
	SubMuli-Wilhelm	-10.48	-21.15	0.19	.06
	Upper-Wilhelm	-21.32	-30.39	-12.24	< .001
Beryllium	SubMuli-Upper	-0.26	-0.51	-0.02	.03
	SubMuli-Wilhelm	0.05	-0.29	0.40	.99
	Upper-Wilhelm	0.31	0.02	0.61	.03
Saturated	SubMuli-Upper	1.90	1.02	2.78	< .001
	SubMuli-Wilhelm	-0.23	-1.47	1.00	.99
	Upper-Wilhelm	-2.14	-3.19	-1.08	< .001
Aromatic	SubMuli-Upper	-0.28	-2.42	1.85	.99
	SubMuli-Wilhelm	-6.06	-9.08	-3.04	< .001
	Upper-Wilhelm	-5.77	-8.34	-3.20	< .001

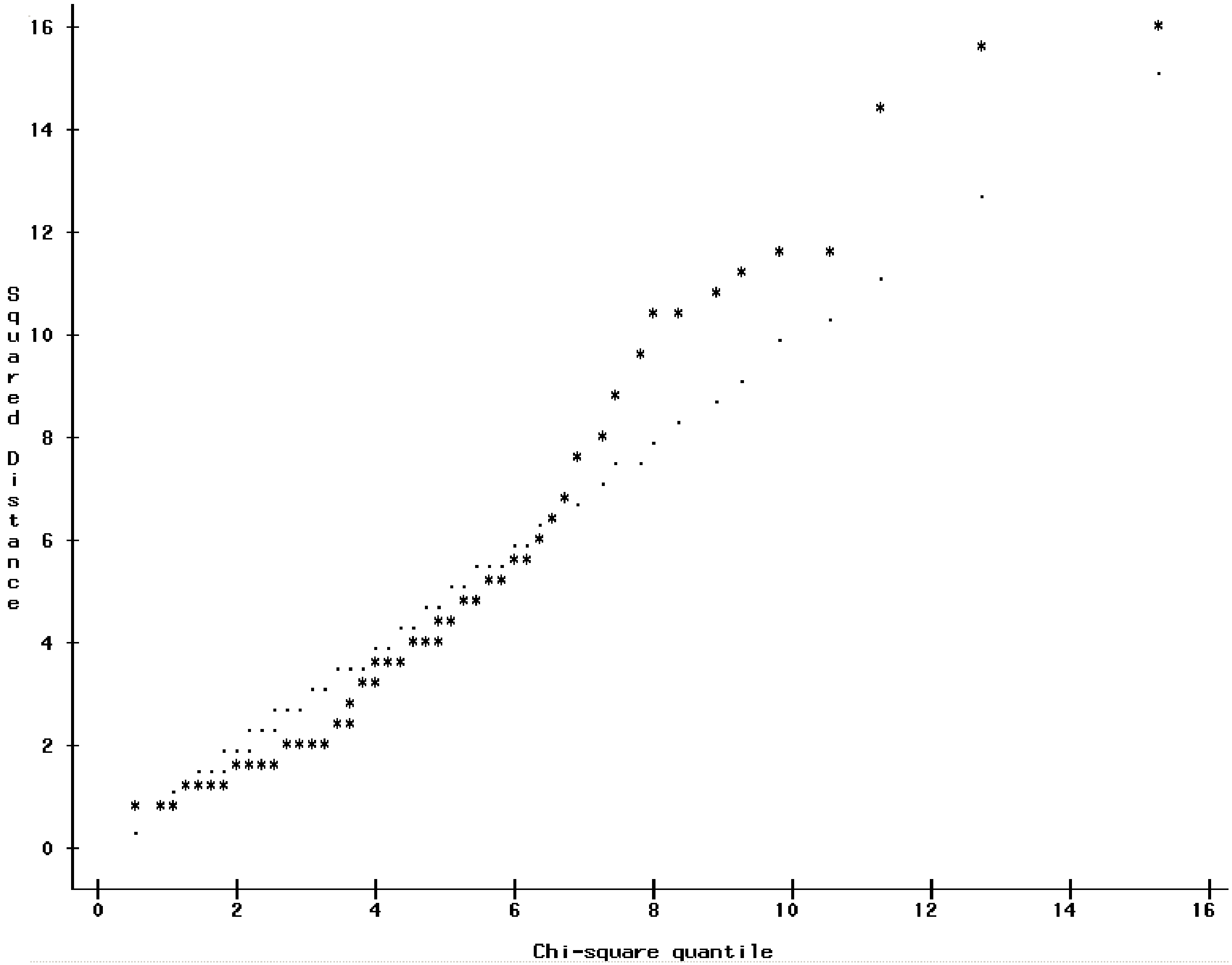
Table 6 shows all of the follow-up multiple comparisons for the three zones for each of the five chemicals. Since the corresponding Bonferroni corrected *p*-values have been adjusted for the number of tests we are conducting we simply need to investigate the mean differences for probability values less than .05. For the first four chemicals (Vanadium, Iron, Beryllium, and Saturated) we observed statistically significant mean differences between crude oil samples taken from *SubMuli* and *Upper* as well as *Upper* and *Wilhelm* zones. However, for these four chemicals there were no significant differences between the means of the *SubMuli* and *Wilhelm* zones. Also, since *Wilhelm* can be seen as similar (non-significantly statistically different) from *SubMuli* it is unsurprising that in those chemical comparisons the *Upper-Wilhelm* comparison is always significant, much like the one between *SubMuli-Upper.*

Unlike the first four chemicals, Aromatic Hydrocarbons do not vary significantly between the *SubMuli* and the *Upper* zones. This means that for this chemical the two groups have approximately equal means. Subsequently, the average Aromatic Hydrocarbon chemical presence in *SubMuli* is comparable to that of the *Upper* zone. Since this is the case, much like observed in the earlier situations, the comparison of Wilhelm against either of these groups is statistically significant.

9) Look at the distribution of residuals (by themselves and grouped by zone). Discuss their corresponding distributions.

By running the MANOVA we also are able to export the individual residuals for each of the five variables. These are the model residuals which inform us about the fit of the overall model. One of the assumptions in MANOVA, much like in the univariate ANOVA analogy, is that the errors should have a multivariate normal distribution. Given that each group is evaluated separately this assumption ought to be investigated much like multivariate normality within the original dataset. Figure 3 shows a pooled evaluation of all of the residuals from the model. Since these groups are independent the pooled residual evaluation is not normal.

Figure 3. Squared distance chi-square Q-Q plot for MANOVA residuals



Each group's residuals should be assessed separately, which would investigate the model fit for that group. Table 7 summarizes the multivariate normality assessment of the residuals for each group. Here we can observe that the residuals for the *Upper* zone, unlike *SubMuli* and *Wilhelm*, are not multivariate normally distributed. Given earlier assessments within the *Upper* zone this is not surprising. The Upper data was skewed and so a deviation from the multivariate normality assumption was to be expected. Further investigation would be needed into the nature of the skew and it's direct effects on the model assessment.

Table 7. Mardia's coefficients for residual distribution by group

Group	Skewness	*p*	Kurtosis	*p*	Henze-Zirkler	*p*
SubMuli	47.03	.08	-.47	.64	2.25	.02
Upper	92.62	< .01	1.88	.06	6.25	< .01
Wilhelm	31.97	.62	-1.49	.14	.76	.45

10) What would be your overall final assessment of the data? What overarching conclusions would you make regarding the siphoning of crude oil and the zone's from which it is taken?

Concerning the chemical composition of Vanadium, Iron, Beryllium, and Saturated Hydrocarbons in crude oil siphoned at the SubMuli and Wilhelm are approximately the same. If the chemical of interest in the crude oil sample taken was any one of these four it would make no difference whether the crude oil would be siphoned at *SubMuli* or *Wilhelm*. However, regarding the content of Aromatic Hydrocarbons in the crud oil is differentiated least between *SubMuli* and *Upper*. It is important to note that even those two

Knowing these mean differences also allows for a classification of two crude oil samples. Given that you are presented with two samples of crude oil you would only have to investigate the amounts of Aromatic Hydrocarbons and any one of the other chemicals, let's say Vanadium. You could then compare the amounts.

Table 8 demonstrates how the comparison would be conducted. Let us say that we compared the Vanadium amounts and found no significant difference. This would mean that the samples came from either *SubMuli* or *Wilhelm* since Vanadium does not differ between those two zones. Now let us say we investigated the amounts of Aromatic Hydrocarbons. If the two samples did differ then we would be able to say that we have a sample from *Wilhelm* (the one with the higher Aromatic Hydrocarbon content) and one from *SubMuli* (with less Aromatic Hydrocarbons).

Table 8. Classification based on chemical comparison

Chemical	Comparison	Group
Vanadium	$\bar{x}_1 = \bar{x}_2$	*SubMuli* or *Wilhelm*
	$\bar{x}_1 \neq \bar{x}_2$	Higher mean from *Upper* Lower mean from *SubMuli* or *Wilhelm*
Aromatic	$\bar{x}_1 = \bar{x}_2$	*SubMuli* or *Upper*
	$\bar{x}_1 \neq \bar{x}_2$	Higher mean from *Wilhelm* Lower mean from *SubMuli* or *Upper*

Extra Credit: *Verify Wilk's Lambda using either SPSS, SAS or Excel*

Computations for the extra credit are attached. You may also investigate the online posted Excel spreadsheet to see the corresponding matrix operations.

```
/* ---------------------------------------------------------- */
/*                  Hmw 4: MANOVA (SAS Syntax)                */
/* ---------------------------------------------------------- */

/* Importing the data into the temporary work folder */
proc import file='E:\Teaching\Mulivariate S09\Lab 7\Homework
     4\crudeoil.xls'
     out = crudeoil
     dbms = EXCEL2000;
     getnames=yes;
data crudeoil;
     set crudeoil;
     id = _N_;
run;

/* QUESTION 1 */

proc corr data=crudeoil;
run;

/* ---------------------------------------------------- */
/* !!! RUN THE PROC IML SCATTER PLOT MATRIX ROUTINE !!! */
/* ---------------------------------------------------- */

/* QUESTION 2 */

proc sort data=crudeoil;
     by zone;
proc means data=crudeoil mean stderr std ;
     by zone;
run;

/* QUESTION 4 */

data submuli;
     set crudeoil;
     if zone = 'SubMuli';
data wilhelm;
     set crudeoil;
     if zone = 'Wilhelm';
data upper;
     set crudeoil;
     if zone = 'Upper';
run;

/* Let us investigate the leverage values for each of our zones
     */
proc reg data=submuli noprint;
     model id = vanadium iron beryllium saturated aromatic /
     influence;
     output out=submuli H=Leverage;
```

```
proc reg data=wilhelm noprint;
     model id = vanadium iron beryllium saturated aromatic /
     influence;
     output out=wilhelm H=Leverage;
proc reg data=upper noprint;
     model id = vanadium iron beryllium saturated aromatic /
     influence;
     output out=upper H=Leverage;
run;
quit;

/* Critical leverage values we should consider */
proc iml;

     Nsub = 11;
     Nupp = 38;
     Nwill = 7;

     crit1_sub = (2*5)/Nsub;
     crit1_upp = (2*5)/Nupp;
     crit1_will = (2*5)/Nwill;
     crit1 = (crit1_sub||crit1_upp||crit1_will)`;

     crit2_sub = 2*((5+1)/Nsub);
     crit2_upp = 2*((5+1)/Nupp);
     crit2_will = 2*((5+1)/Nwill);
     crit2 = (crit2_sub||crit2_upp||crit2_will)`;

     crit3_sub = ((2*(gaminv(.99,2.5)))/(Nsub-1))+(1/Nsub);
     crit3_upp = ((2*(gaminv(.99,2.5)))/(Nupp-1))+(1/Nupp);
     crit3_will = ((2*(gaminv(.99,2.5)))/(Nwill-1))+(1/Nwill);
     crit3 = (crit3_sub||crit3_upp||crit3_will)`;

     zone = {'SubMuli','Upper','Wilhelm'};

     print 'Leverage cut-off values';
     print zone crit1 crit2 crit3;

quit;

proc means data=submuli max;
     var Leverage;
proc means data=upper max;
     var Leverage;
proc means data=wilhelm max;
     var Leverage;
run;

/* ------------------------------ */
/* !!! RUN THE MULTNORM2 MACRO !!! */
/* ------------------------------ */
/* Let's assess multivariate normality */
%multnormplt (data=submuli,
     var= vanadium iron beryllium saturated aromatic,
     title='Submuli Zone');
%multnormplt (data=upper,
     var= vanadium iron beryllium saturated aromatic,
     title='Upper Zone');
%multnormplt (data=wilhelm,
     var= vanadium iron beryllium saturated aromatic,
     title='Wilhelm Zone');
quit;

/* Bartlett's Test */
proc discrim data=crudeoil pool=test;
  class zone;
  var vanadium iron beryllium saturated aromatic;
run;

/* QUESTION 6 */

/* Running the MANOVA and saving out residuals */
proc glm data=crudeoil;
  class zone;
  model vanadium iron beryllium saturated aromatic = zone;
  lsmeans zone / stderr cl pdiff adjust=Bon;
  manova h=zone / printe printh;
  output out=resids r=rva rir rbe rsa rar;
run;
quit;

/* Let us create the corresponding profile plot */
title "Profile Plot for Crude Oil Data";
data crudflat;
  set crudeoil;
  chemical="vanadium"; amount=vanadium; output;
  chemical="iron"; amount=iron; output;
  chemical="beryllium"; amount=beryllium; output;
  chemical="saturated"; amount=saturated; output;
  chemical="aromatic"; amount=aromatic; output;
  keep zone chemical amount;
proc sort;
  by zone chemical;
proc means noprint;
  by zone chemical;
  var amount;
  output out=chemmeans mean=mean;
run;

proc gplot data=chemmeans;
  axis1 length=4.5 in;
  axis2 length=7.5 in;
  plot mean*chemical=zone / vaxis=axis1 haxis=axis2;
  symbol1 v=J f=special h=2 l=1 i=join color=black;
  symbol2 v=K f=special h=2 l=1 i=join color=black;
  symbol3 v=L f=special h=2 l=1 i=join color=black;
  symbol4 v=M f=special h=2 l=1 i=join color=black;
run;
quit;

/* QUESTION 9 */

proc univariate data=resids plot;
     var rva rir rbe rsa rar;
     by zone;
run;
quit;

/* Run the multnorm2 SAS macro */
data submuli;
     set resids;
     if zone = 'SubMuli';
data upper;
     set resids;
     if zone = 'Upper';
data Wilhelm;
     set resids;
     if zone = 'Wilhelm';
run;

%multnormplt (data=submuli,
     var= rva rir rbe rsa rar,
     title="Crude oil data residuals");
%multnormplt (data=upper,
     var= rva rir rbe rsa rar,
     title="Crude oil data residuals");
%multnormplt (data=wilhelm,
     var= rva rir rbe rsa rar,
     title="Crude oil data residuals");
quit;
```