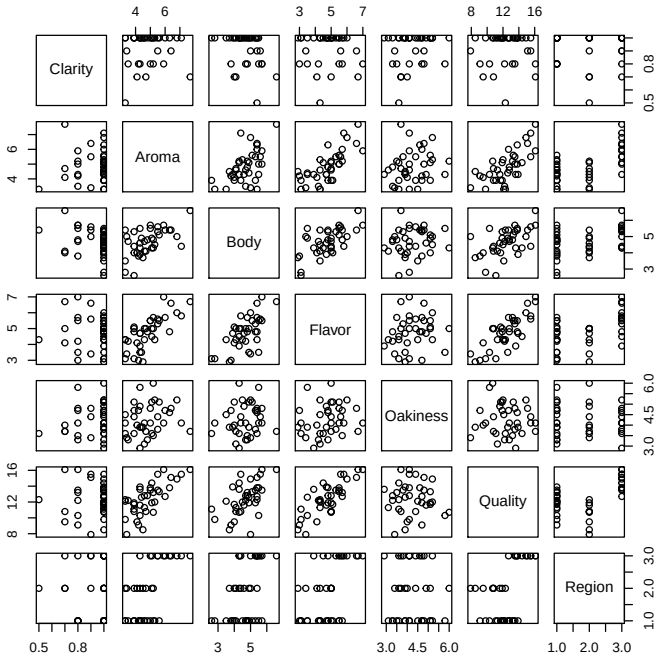
Question 2

A: Fit a linear regression model relating wine quality to the first 5 regressors except Region.

```
> library(xtable)
> data<-read.csv("wine_data.csv")
> pairs(data)
```



As we can see Region is quite clearly a categorical variable while the others are continuous. Quality seems to be linear with the other continous variables.

```
> fit<-lm(Quality~Oakiness+Flavor+Body+Aroma+Clarity,data=data)
> print(xtable(summary(fit)))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.9969	2.2318	1.79	0.0828
Oakiness	-0.6840	0.2712	-2.52	0.0168
Flavor	1.1683	0.3045	3.84	0.0006
Body	0.2732	0.3326	0.82	0.4175
Aroma	0.4826	0.2724	1.77	0.0861
Clarity	2.3395	1.7348	1.35	0.1870

Here is the model for our wine quality. The equation becomes Quality=4-.684*Oakiness+1.1683*Flavour+0.2732*Body+0.4826*Aroma+2.3395*Clarity

B: Construct an Anova table and test for significance of regression

```
> fit_0<-lm(Quality~1,data=data)
> print(xtable(anova(fit_0,fit)))
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	37	154.79				
2	32	43.25	5	111.54	16.51	0.0000

C:

```
> print(xtable(summary(fit)))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.9969	2.2318	1.79	0.0828
Oakiness	-0.6840	0.2712	-2.52	0.0168
Flavor	1.1683	0.3045	3.84	0.0006
Body	0.2732	0.3326	0.82	0.4175
Aroma	0.4826	0.2724	1.77	0.0861
Clarity	2.3395	1.7348	1.35	0.1870

At a significance level of 0.05. We can say that Body and Clarity are not statistically significant. While Aroma, Flavour and Oakiness are statistically significant, we have enough evidence to reject the null hypothesis in their cases.

D:

```
> fit_2<-lm(Quality~Aroma+Flavor+Oakiness,data=data)
> full_model<-c(summary(fit)$r.squared,summary(fit)$adj.r.squared)
> reduced_model<-c(summary(fit_2)$r.squared,summary(fit_2)$adj.r.squared)
> r_table<-cbind(full_model,reduced_model)
> colnames(r_table)<-c("Full Model","Reduced Model")
> rownames(r_table)<-c("R-Squared","Adj R-Squared")
> print(xtable(r_table))
```

	Full Model	Reduced Model
R-Squared	0.72	0.70
Adj R-Squared	0.68	0.68

As we can see removing the insifignificant variables did not affect our mode significantly the R-squared values did not decease by much, illustrating that the variables we removed were not contributing much to the explanatory power of our model.

E: Partial F-Test