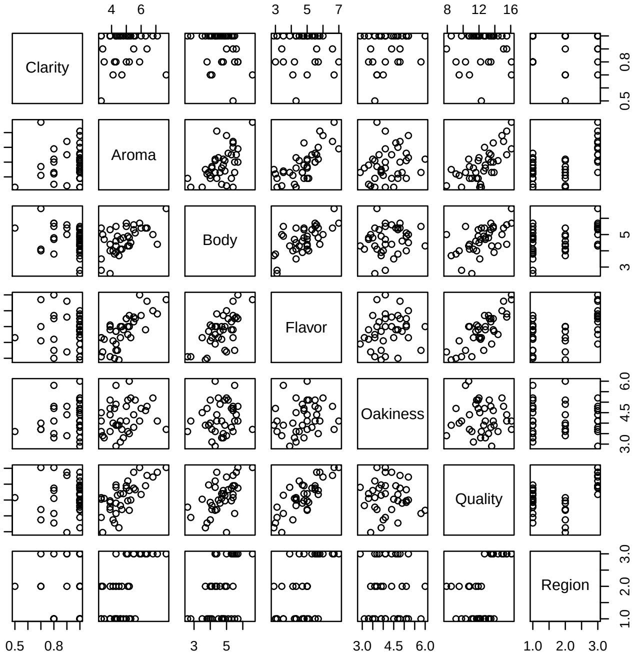
Question 2

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

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

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
Clarity	2.3395	1.7348	1.35	0.1870
Aroma	0.4826	0.2724	1.77	0.0861
Body	0.2732	0.3326	0.82	0.4175
Flavor	1.1683	0.3045	3.84	0.0006
Oakiness	-0.6840	0.2712	-2.52	0.0168

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

D:

```
> fit_2<-lm(Quality~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	0.66
Adj R-Squared	0.68	0.64

As we can see removing the insifignificant variables did not affect our model 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

To test the hypothesis we create a reduced model containing only B4,B5 which are Oakiness and Flavor.We then do an F test to check wether the variables that we removed were adding to the explanatory power of the model.

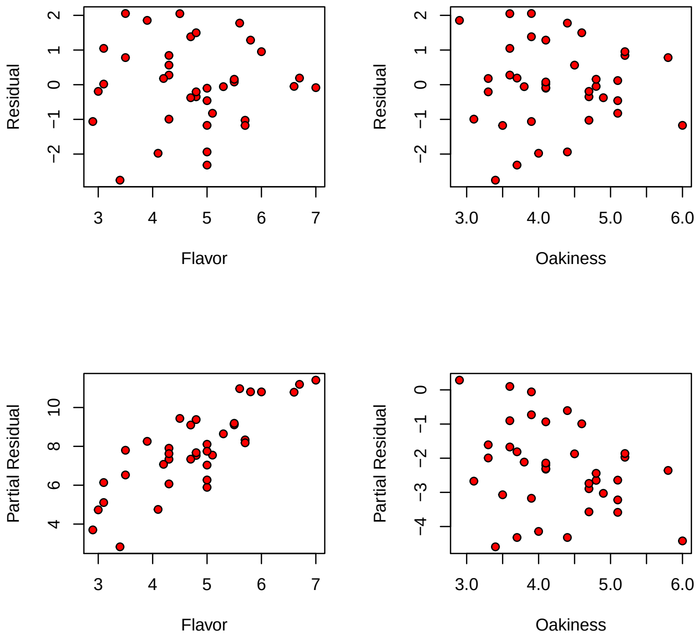
```
> full_model<-lm(Quality~Clarity+Aroma+Body+Flavor+Oakiness,data=data)
> reduced_model<-lm(Quality~Flavor+Oakiness,data=data)
> print(xtable(anova(reduced_model,full_model)))


>
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	35	52.46				
2	32	43.25	3	9.21	2.27	0.0992

From the P-value of the resulting F test, we cannot reject the null hypothesis at a 5 percent level of significance. As such we conclude that B1=B2=B3=0 and we have no reason for keeping them in our model. For further analysis we will utilize the reduced model.

F: Residual Analysis

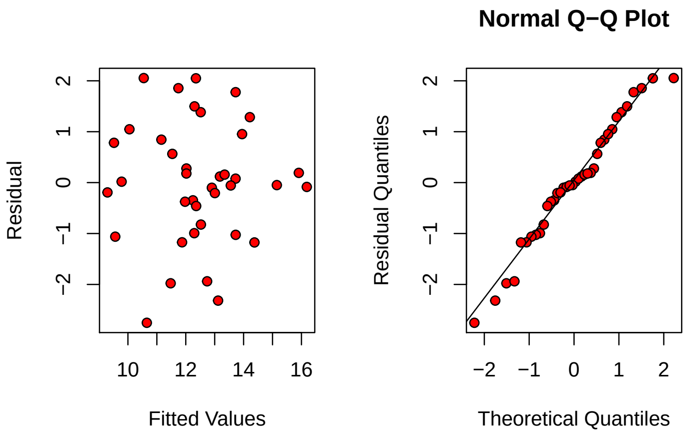
```
> attach(data)
> res.lm1<-residuals(reduced_model)
> pres_model_Flav<-res.lm1+Flavor*coef(reduced_model)[2]
> pres_lm1_Oak <- res.lm1 + Oakiness*coef(reduced_model)[3]
> par(mfrow = c(2,2))
> plot(Flavor,res.lm1,pch = 21, bg='red',xlab='Flavor',ylab ='Residual')
> plot(Oakiness,res.lm1,pch = 21, bg='red',xlab='Oakiness',ylab ='Residual')
> plot(Flavor,pres_model_Flav,pch = 21, bg='red',xlab='Flavor',ylab ='Partial Residual')
> plot(Oakiness,pres_lm1_Oak,pch = 21, bg='red',xlab='Oakiness',ylab ='Partial Residual')
> detach(data)
```



For the Variable Flavor, the Residual vs Flavor plot does indeed show random scatter about 0, and the corresponding Partial Residual PLot doe have a linear relationship. We can say with some considendce that the expectation of residuals is indeed 0.

For the variable Oakiness, the scatter is less random as there is considerably fewer point as oakiness grows larger. In addition, the partial residuals plot does not exhibit a nice linear relationship. We can't say with great confidence that the expectation of residuals is equal to 0.

```
> par(mfrow = c(1,2))
> haty.lm1 = predict(reduced_model)
> plot(haty.lm1,res.lm1,pch = 21, bg='red',xlab='Fitted Values',ylab ='Residual')
> qqnorm(res.lm1,pch = 21, bg='red',ylab ='Residual Quantiles')
> qqline(res.lm1)
>
```



From the Residual vs Fitted plot we can see that the variance is relatively constant throughout, except for the Fitted Range 15-16 in which the residuals exhibit a much smaller spread. Due to the majority of the interval the spread is constant, the constant variance assumption is satisfied.

From the normal-Q-Q plot we can see that the Residual Quantiles are clustered around the line as such the residuals do indeed have a normal distribution.

Without knowing the manner in which this data was collected we cannot comment on the independence of errors assumption.

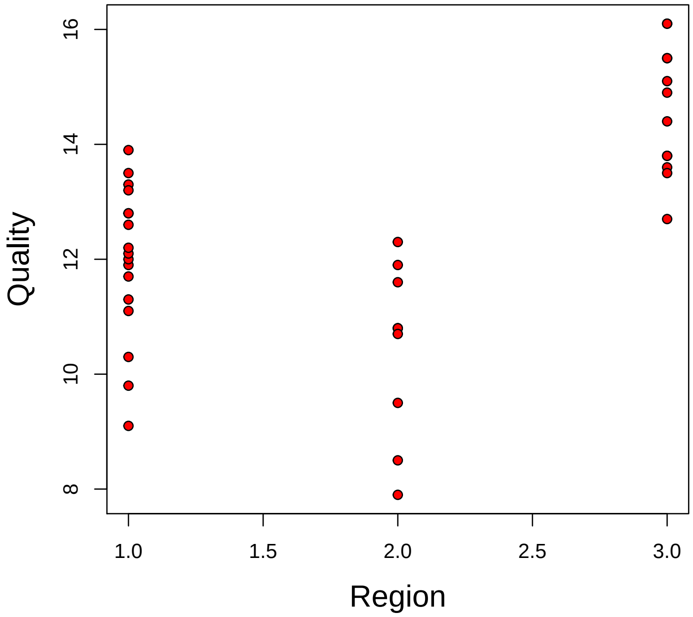
G:

```
> conf<-predict(reduced_model,newdata=data.frame(Flavor=5.0,Oakiness=5.5),
+                                 interval=("confidence"),level=0.95)
> colnames(conf)<-c("Prediction","Lower","Upper")
> print(xtable(conf))
```

	Prediction	Lower	Upper
1	12.14	11.34	12.94

H:

```
> attach(data)
> plot(Region,Quality,pch = 21, bg='red',ylab='Quality',xlab='Region',cex.lab=1.5)
> detach(data)
```



I took region 2 to be the base case as it was the lowest.

```
> region<-data$Region
> fac<-factor(region)
> region<-relevel(fac,'2')
> fit_r<-lm(Quality~region,data=data)
> print(xtable(summary(fit_r)))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	10.4444	0.4371	23.90	0.0000
region1	1.5320	0.5405	2.83	0.0076
region3	4.1389	0.5782	7.16	0.0000

As we can see going from region 2 to region 1 increases wine quality by 1.5320, and going from region 2 to region 3 we see an increase of wine quality of 4.1389. Both slopes are statistically significant.

```
> print(xtable(anova(fit_r)))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
region	2	94.62	47.31	27.52	0.0000
Residuals	35	60.17	1.72		

As we can see the p-value fromt he resulting anova test is effectively zero. Thus, we have strong evidence that the means of quality are different for at least 2 of the 3 regions.