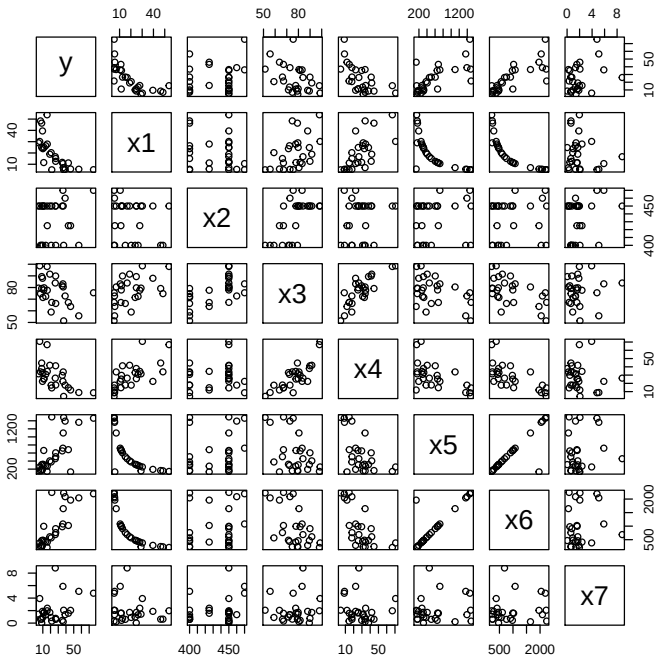
Question 2

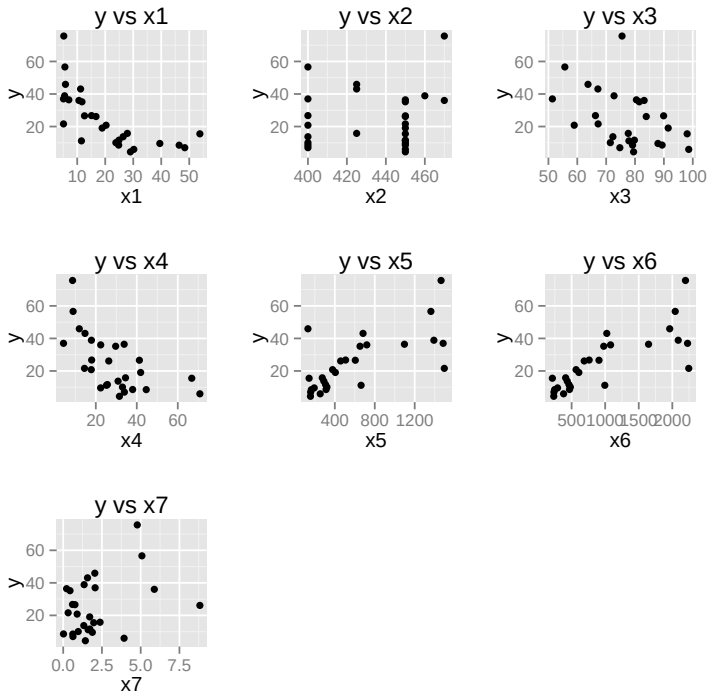
A: Plot the response y vs each of the covariates.

```
> df<-read.csv("Belle_data.csv")
> rownames(df) <- df[, 1] ## set rownames
> df <- df[, -1]
> pairs(df)
> df2<-df
> colnames(df)<-c("CO2","SpaceTime","Temp","PercentSolvation","OilYield","CoalTotal","Solver
>
```



Here we see a very messy pairs plot. As this data set has many variables and we are interested in the relationship between y and the x variables and not necessarily between x variables themselves.

```
> library(ggplot2);library(gridExtra);library(xtable);
> plot1<-qplot(x1,y,data=df2)+ggtitle("y vs x1")
> plot2<-qplot(x2,y,data=df2)+ggtitle("y vs x2")
> plot3<-qplot(x3,y,data=df2)+ggtitle("y vs x3")
> plot4<-qplot(x4,y,data=df2)+ggtitle("y vs x4")
> plot5<-qplot(x5,y,data=df2)+ggtitle("y vs x5")
> plot6<-qplot(x6,y,data=df2)+ggtitle("y vs x6")
> plot7<-qplot(x7,y,data=df2)+ggtitle("y vs x7")
> grid.arrange(plot1,plot2,plot3,plot4,plot5,plot6,plot7,ncol=3)
```



From these plot we can clearly see that x1,x2,x3,x4 have a seemingly negative relationship with y. While x5,x6 and x7 have a positive correlation.

B: Fit Multiple regrresion model relating CO2 product to total solvent and hydrogen consumption.

```
> fit<-lm(CO2~SolventTotal+HydrogenConsumption,data=df)
> print(xtable(summary(fit)))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.5265	3.6101	0.70	0.4908
SolventTotal	0.0185	0.0027	6.74	0.0000
HydrogenConsumption	2.1858	0.9727	2.25	0.0341

C: Test the significance of regression and calculate R2.

```
> f <- summary(fit)$fstatistic
> prob <- round(pf(f[1],f[2],f[3],lower.tail=F),digits=10)
> print(xtable(anova(fit)))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
SolventTotal	1	5008.94	5008.94	50.86	0.0000
HydrogenConsumption	1	497.34	497.34	5.05	0.0341
Residuals	24	2363.84	98.49		

The F test of significanse yields an F Statisitc of 27.95259. with a corresponding p value of 5.4e-07. The R-squared value comes out to 0.69964. Clearly there is at least one variable in our regreesion that has some explanatory power.

D: Now do the test in (c) using t tests and determine the contribution of the two regressors.

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

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.5265	3.6101	0.70	0.4908
SolventTotal	0.0185	0.0027	6.74	0.0000
HydrogenConsumption	2.1858	0.9727	2.25	0.0341

Both are significant. Funnily enough the intercept is not significantly different from zero.

E: Construct the 95 percent CI on the regression parameters in (b).

```
> print(xtable(confint(fit)))
```

	2.5 %	97.5 %
(Intercept)	-4.92	9.98
SolventTotal	0.01	0.02
HydrogenConsumption	0.18	4.19

F: Reconstruct the overall ANOVA table using the summary results.

```
> source1<-c("SolventTotal","HydrogenConsumpetion","Residuals")
> degf<-c(1,1,length(df$CO2)-3)
> dfr <- df.residual(fit)
> p <- fit$rank
> p1 <- 1L:p
> ssr <- sum(fit$residuals^2)
> comp <- fit$effects[p1]
> asgn <- fit$assign[fit$qr$pivot][p1]
> ss <- c(unlist(lapply(split(comp^2, asgn), sum)), ssr)
> ss1<-ss[2:4]
> ms<-ss1/degf
> f <- round(ms/(ssr/dfr),3)
> f[3]<-0
> P <- round(pf(f, degf, dfr, lower.tail = FALSE),3)
> P[3]<-0
> data_tab<-cbind(source1,degf,ss1,ms,f,P)
> row.names(data_tab)<-NULL
> rownames(data_tab) <- data_tab[, 1]
> data_tab <- data_tab[, -1]
> print(xtable(data_tab))
```

	degf	ss1	ms	f	P
SolventTotal	1	5008.93619257035	5008.93619257035	50.856	0
HydrogenConsumpetion	1	497.340747084587	497.340747084587	5.049	0.034
Residuals	24	2363.83515664136	98.4931315267233	0	0

For comparison, on the next page is the Anova table that R produces.

```
> print(xtable(anova(fit)))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
SolventTotal	1	5008.94	5008.94	50.86	0.0000
HydrogenConsumption	1	497.34	497.34	5.05	0.0341
Residuals	24	2363.84	98.49		

As we can see both the tables are identical.