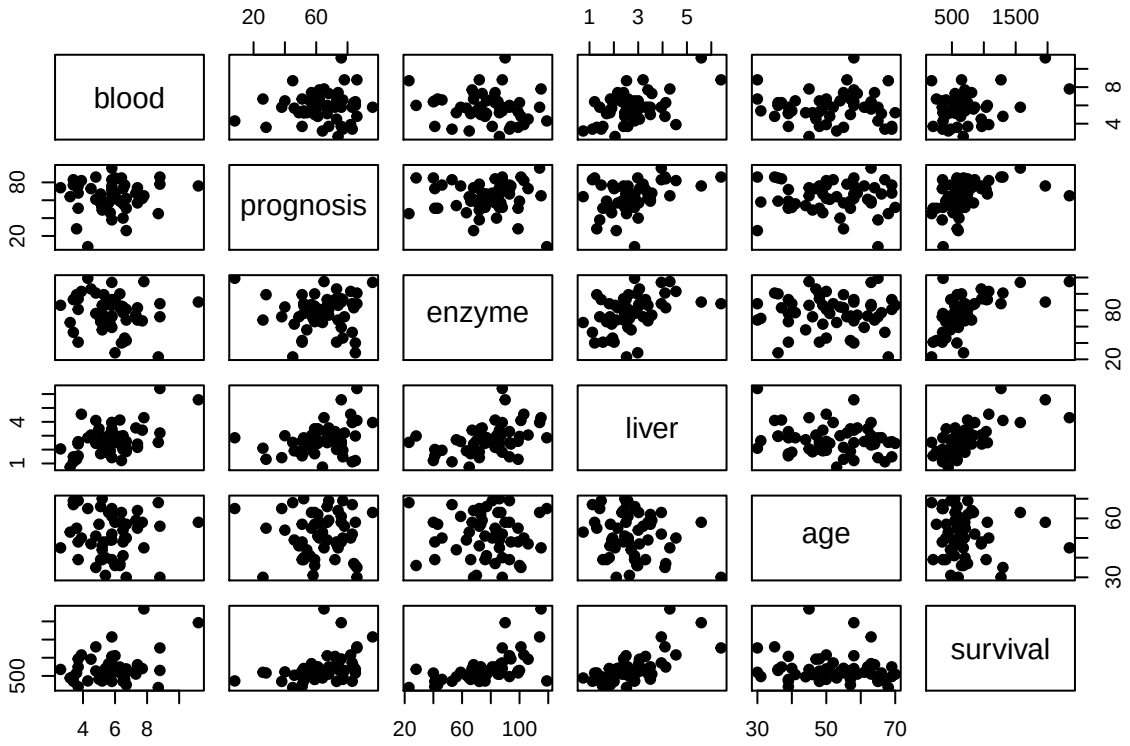
Assignment Umair Amjad 46462252

#q1a)

```
mr_dat = read.table("surg.dat", header = TRUE)
head(mr_dat)
```

```
##   blood prognosis enzyme liver age gender survival
## 1   6.7        62     81  2.59  50      M      695
## 2   5.1        59     66  1.70  39      M      403
## 3   7.4        57     83  2.16  55      M      710
## 4   6.5        73     41  2.01  48      M      349
## 5   7.8        65    115  4.30  45      M     2343
## 6   5.8        38     72  1.42  65      F      348
```

```
plot(mr_dat[,-6], pch=19)
```



the gender variable needs to be removed as gender is a categorical variable and it can alter the results of the scatterplot significantly the plot sugests that the dots are significantly close to each suggesting that the data isnt evenly spread.

By analysisng the matrix we can conclude that survival has a stronger relationship with enyzme and liver. Survival has moderate relationship with blood and week relationship with age

q1b)

```
mr_dat = read.table("surg.dat", header = TRUE)
round(cor(mr_dat[,-6]),2)
```

```
##           blood prognosis enzyme liver   age survival
## blood      1.00      0.09  -0.15  0.50 -0.02     0.35
## prognosis  0.09      1.00  -0.02  0.37 -0.05     0.42
## enzyme    -0.15     -0.02   1.00  0.42 -0.01     0.58
## liver      0.50      0.37   0.42  1.00 -0.21     0.67
## age       -0.02     -0.05  -0.01 -0.21  1.00    -0.12
## survival   0.35      0.42   0.58  0.67 -0.12     1.00
```

Therefore we can see liver has a better relationship with the survival variable while on the other hand, the worst relationship is between the variables age and enyzme

q1c)

```
mr_dat = read.table("surg.dat", header = TRUE)
mylm = lm(survival~. , data = mr_dat)
summary(mylm)
```

```
##
## Call:
## lm(formula = survival ~ ., data = mr_dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -388.25 -147.61   11.72  124.67  954.44
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1179.1889   283.8232  -4.155 0.000136 ***
## blood          86.6437    27.4920   3.152 0.002825 **
## prognosis       8.5013     2.1601   3.936 0.000273 ***
## enzyme         11.1246     1.9820   5.613 1.03e-06 ***
## liver          38.5068    51.7967   0.743 0.460926
## age            -2.3409     3.0141  -0.777 0.441257
## genderM        -0.2201    67.5146  -0.003 0.997413
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 233.1 on 47 degrees of freedom
## Multiple R-squared:  0.695,  Adjusted R-squared:  0.656
## F-statistic: 17.85 on 6 and 47 DF,  p-value: 1.19e-10
```

```
anova(mylm)
```

```
## Analysis of Variance Table
##
## Response: survival
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## blood      1 1005152 1005152 18.5060 8.502e-05 ***
## prognosis  1 1278496 1278496 23.5385 1.387e-05 ***
## enzyme     1 3442172 3442172 63.3742 2.915e-10 ***
## liver      1   57862   57862  1.0653    0.3073
## age        1   33032   33032  0.6082    0.4394
## gender     1       1       1  0.0000    0.9974
## Residuals 47 2552807   54315
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
mylm8= lm(survival~ blood+prognosis+ enzyme, data= mr_dat)
anova(mylm8)
```

```
## Analysis of Variance Table
##
## Response: survival
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## blood      1 1005152 1005152  19.010 6.484e-05 ***
## prognosis  1 1278496 1278496  24.180 9.883e-06 ***
## enzyme     1 3442172 3442172  65.101 1.303e-10 ***
## Residuals 50 2643701   52874
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

survival = -11.367 + 86.630(blood)+8.501(prognosis)+11.125(Enzyme)+38.507(Liver)-2.340(age)-0.2201(gender)

$H_0: _1 = _2……. _k$ $H_1 : \beta_i \neq 0$

therefore due to the pvalue being 1.19e-10 therefore we have evidence to reject this null hypothesis as 1.19e-10<0.05 and the f statistic is 17.85 with the degree ofs freedom between 6 and 47. The anova table suggests that the pvalue of blood,prognesis,enyzme is less then 0.05 while on the other hand the pvalues of liver,age and gender is greater then 0.05 therefore these variables shouldnt be included in the equation. for the anova table liver,age,gender pvalue >0.05 meaning that these variable needs to be remove from the anova table and once the variable were remove from the anova table blood pvalue increase, enyzme pvalue increase while the enyzme pvalue got smaller. therefore is a relationship between aall the predictors

q1d)

```
mylm2 = lm(survival~ blood+prognosis+enzyme+liver+age, data= mr_dat)
summary(mylm2)
```

```
##
## Call:
```

```
## lm(formula = survival ~ blood + prognosis + enzyme + liver +
##     age, data = mr_dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -388.34 -147.74   11.74  124.67  954.32
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1179.367    275.619  -4.279 8.91e-05 ***
## blood          86.630     26.905   3.220 0.002302 **
## prognosis       8.501      2.137   3.978 0.000234 ***
## enzyme         11.124      1.958   5.683 7.62e-07 ***
## liver          38.554     49.251   0.783 0.437595
## age            -2.340      2.969  -0.788 0.434514
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 230.6 on 48 degrees of freedom
## Multiple R-squared:  0.695,  Adjusted R-squared:  0.6632
## F-statistic: 21.87 on 5 and 48 DF,  p-value: 2.386e-11
```

```
mylm3 = lm(survival~ blood+prognosis+enzyme+age, data= mr_dat)
summary(mylm3)
```

```
##
## Call:
## lm(formula = survival ~ blood + prognosis + enzyme + age, data = mr_dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -416.92 -142.56  -13.98  138.10  943.31
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1246.655    260.835  -4.779 1.64e-05 ***
## blood         100.660     19.987   5.036 6.83e-06 ***
## prognosis       9.291      1.876   4.951 9.14e-06 ***
## enzyme         12.101      1.502   8.058 1.56e-10 ***
## age            -2.986      2.841  -1.051    0.298
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 229.7 on 49 degrees of freedom
## Multiple R-squared:  0.6911, Adjusted R-squared:  0.6659
## F-statistic: 27.41 on 4 and 49 DF,  p-value: 5.68e-12
```

final model

```
mylm4= lm(survival~ blood+prognosis+enzyme, data = mr_dat)
summary(mylm4)
```

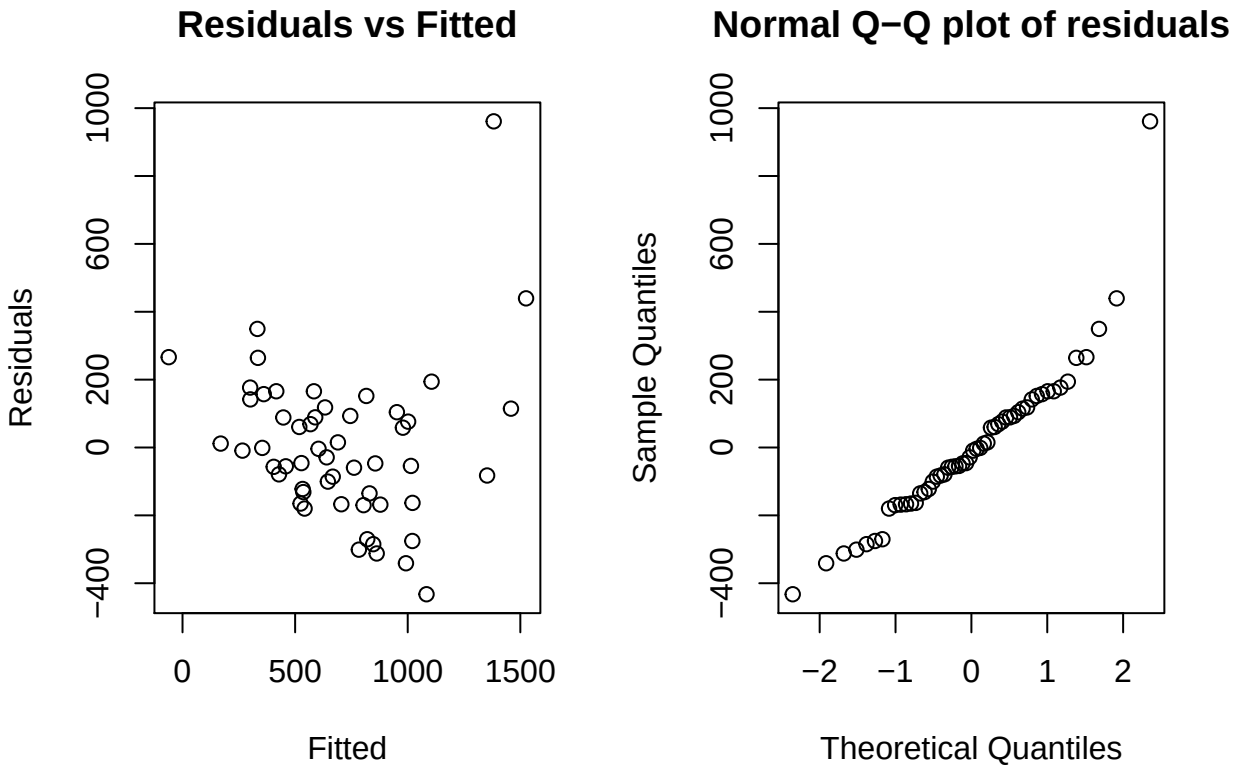
```
##
```

```
## Call:
## lm(formula = survival ~ blood + prognosis + enzyme, data = mr_dat)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -432.4 -134.3  -19.1  111.9  961.1
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1410.847    209.118  -6.747 1.50e-08 ***
## blood         101.054     20.005   5.052 6.22e-06 ***
## prognosis       9.382      1.876   5.000 7.43e-06 ***
## enzyme         12.128      1.503   8.069 1.30e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 229.9 on 50 degrees of freedom
## Multiple R-squared:  0.6841, Adjusted R-squared:  0.6652
## F-statistic:  36.1 on 3 and 50 DF,  p-value: 1.469e-12
```

this model demonstrated above is the best model as the variable gender isnt included due to how gender had a big influences on the model as the summary lm show that liver,age gender needs to be remove from the model as it exceeds the 0.05 significant level so therefore the equation should include survival, enzyme,blood,prognosis survival = -1410.847 +101.054 (blood)+ 9.382(prognosis)+ 12.128(Enzyme)

#q1 e

```
par( mfrow = c(1,2))
mylm6<-lm(survival~ blood+prognosis+enzyme, data=mr_dat)
plot(mylm6$fitted, mylm6$residuals, main = "Residuals vs Fitted",
xlab = "Fitted", ylab = "Residuals")
qqnorm(mylm6$residuals, main = "Normal Q-Q plot of residuals")
```



the normal q-q plots seems to have a close linear action demonstrating that there are errors as the datas gets close to normally distributed. The residuals vs fitted doesnt have a pattern.So therefore the regression model isnt approate model to this case due to these reasons

q1 f

```
mylm5 = lm(log(survival)~ blood+prognosis+enzyme+liver+age , data = mr_dat)
summary(mylm5)
```

```
##
## Call:
## lm(formula = log(survival) ~ blood + prognosis + enzyme + liver +
##     age, data = mr_dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3894 -0.1895  0.0045  0.1782  0.5103
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.047579   0.296655  13.644  < 2e-16 ***
## blood        0.090874   0.028958   3.138  0.00291 **
## prognosis    0.012975   0.002300   5.641 8.82e-07 ***
```

```
## enzyme       0.016126   0.002107   7.654 7.38e-10 ***
## liver        0.010914   0.053010   0.206  0.83775
## age         -0.004584   0.003196  -1.434  0.15796
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2482 on 48 degrees of freedom
## Multiple R-squared:  0.769,  Adjusted R-squared:  0.745
## F-statistic: 31.97 on 5 and 48 DF,  p-value: 3.478e-14
```

```
mylm7= lm(log(survival)~ blood+prognosis+enzyme+age, data = mr_dat)
summary(mylm7)
```

```
##
## Call:
## lm(formula = log(survival) ~ blood + prognosis + enzyme + age,
##     data = mr_dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.39491 -0.18866 -0.00045  0.17491  0.51787
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.028531   0.279090  14.434  < 2e-16 ***
## blood        0.094845   0.021386   4.435 5.20e-05 ***
## prognosis    0.013199   0.002008   6.574 3.04e-08 ***
## enzyme       0.016402   0.001607  10.208 1.01e-13 ***
## age         -0.004767   0.003040  -1.568    0.123
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2458 on 49 degrees of freedom
## Multiple R-squared:  0.7688, Adjusted R-squared:   0.75
## F-statistic: 40.74 on 4 and 49 DF,  p-value: 5.171e-15
```

```
mylm8= lm(log(survival)~ blood+prognosis+enzyme, data = mr_dat)
summary(mylm8)
```

```
##
## Call:
## lm(formula = log(survival) ~ blood + prognosis + enzyme, data = mr_dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.46994 -0.17938 -0.03116  0.17959  0.59105
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.766441   0.226757  16.610  < 2e-16 ***
## blood       0.095475   0.021692   4.401 5.66e-05 ***
## prognosis   0.013344   0.002035   6.558 2.95e-08 ***
## enzyme      0.016444   0.001630  10.089 1.19e-13 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2493 on 50 degrees of freedom
## Multiple R-squared:  0.7572, Adjusted R-squared:  0.7427
## F-statistic: 51.99 on 3 and 50 DF,  p-value: 2.137e-15
```

#q1 g

If the regression model isnt approiate in terms of survival then the next best option is to utilized log therefore log is the best option for survival response. The log method removes all the outliers and make relations more linear

#Q2 a

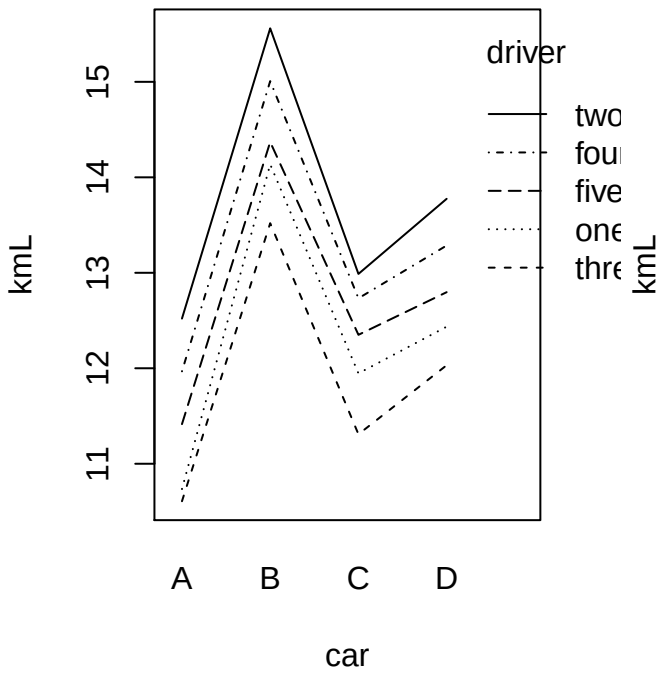
```
car_data = read.table("kml.dat", header=TRUE)
 table(car_data[,c("car","driver")])
```

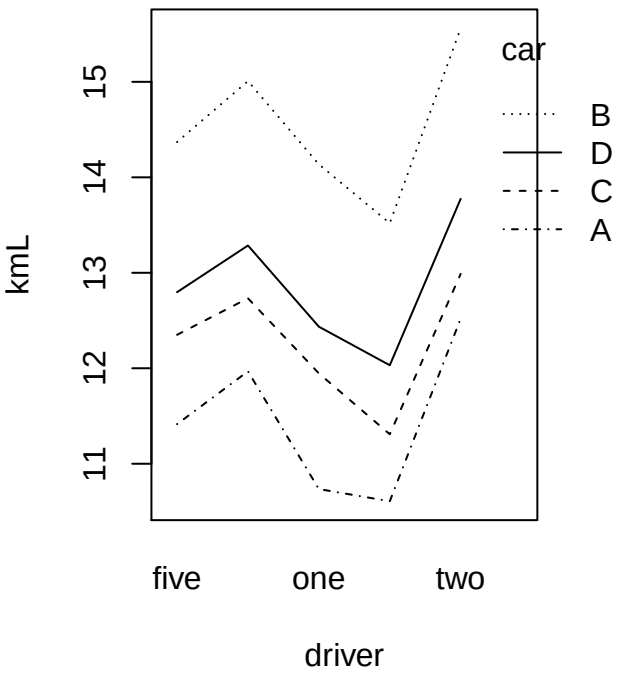
```
##        driver
## car     A B C D
##   five  2 2 2 2
##   four  2 2 2 2
##   one   2 2 2 2
##   three 2 2 2 2
##   two   2 2 2 2
```

This data seems to be balanced due to how it shows the equals number of observations for each variable

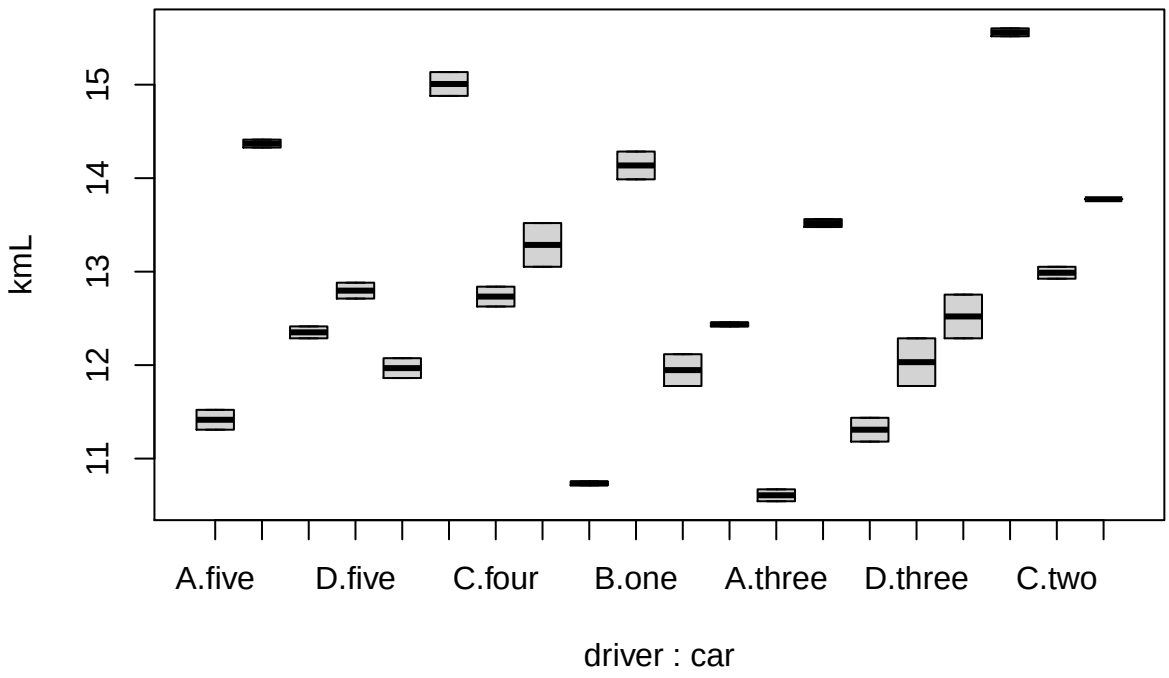
#Q2b

```
 par(mfrow =  c(1,2))
 with(car_data,interaction.plot(driver,car,kmL,trace.label = "driver", xlab= "car", ylab = "kmL"))
 with(car_data,interaction.plot(car,driver,kmL,trace.label="car",xlab = "driver", ylab= "kmL"))
```






```
boxplot(kmL ~ driver + car, data = car_data)
```



Also the prelimary investigation suggest that the lines are parrall meaning that therefore there is no interaction between the lines meaning that the boxplot can be easily intrepret. Therefore the interecation seems to be insignificant.

#Q2c

```
summary(aov(kmL~  driver* car, data=car_data))
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## driver       3  50.66  16.887  531.60  < 2e-16 ***
## car          4  17.12   4.280  134.73 3.66e-14 ***
## driver:car  12   0.44   0.037    1.16    0.371
## Residuals   20   0.64   0.032
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

H_0: $\_1 = \_2 \ldots\ldots \_k$ $H_1 : \beta_i \neq 0$

Through the anova table results we can conclude that the pvalues results is less then 0.05 therefore these results are insignifcant. The assumption in this case can be concluded that there is no interectation between the data as considered before the lines dont intersect, and the pvalues is less then 0.05

#2d The preliminary investigation suggests that the lines are parrall meaning that the interecation is insignifcant this can conclude that the anova results conclude that the pvalues for car and driver is less then 0.05.Therefore those variables shouldnt be removed for this case.