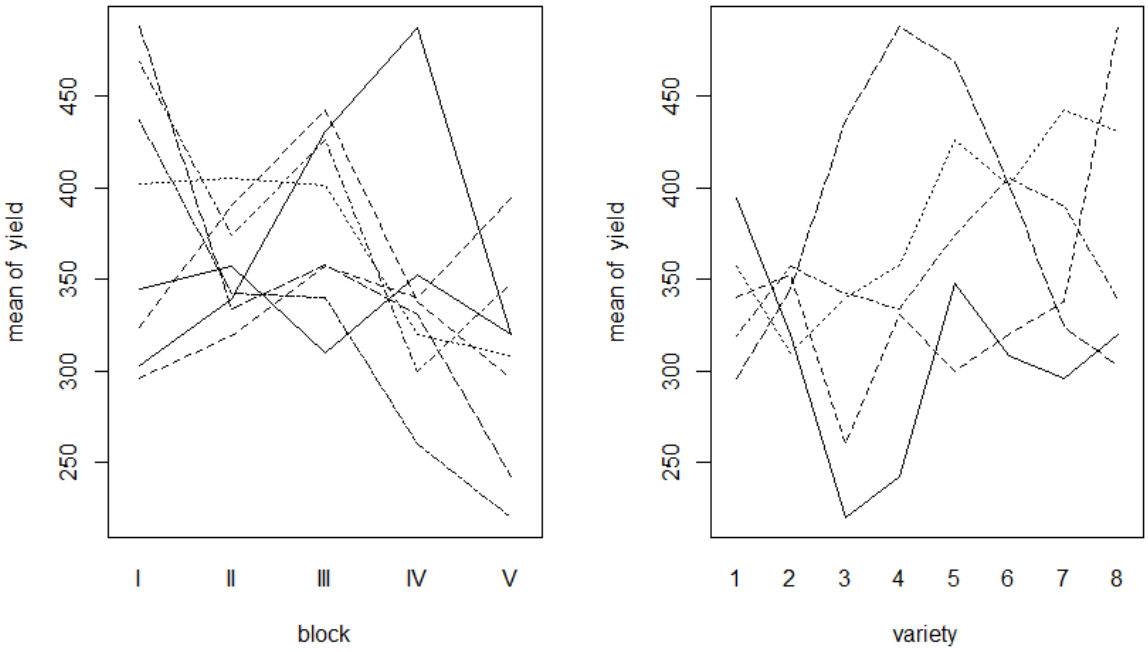
1.

(a) Two-way ANOVA

(b)



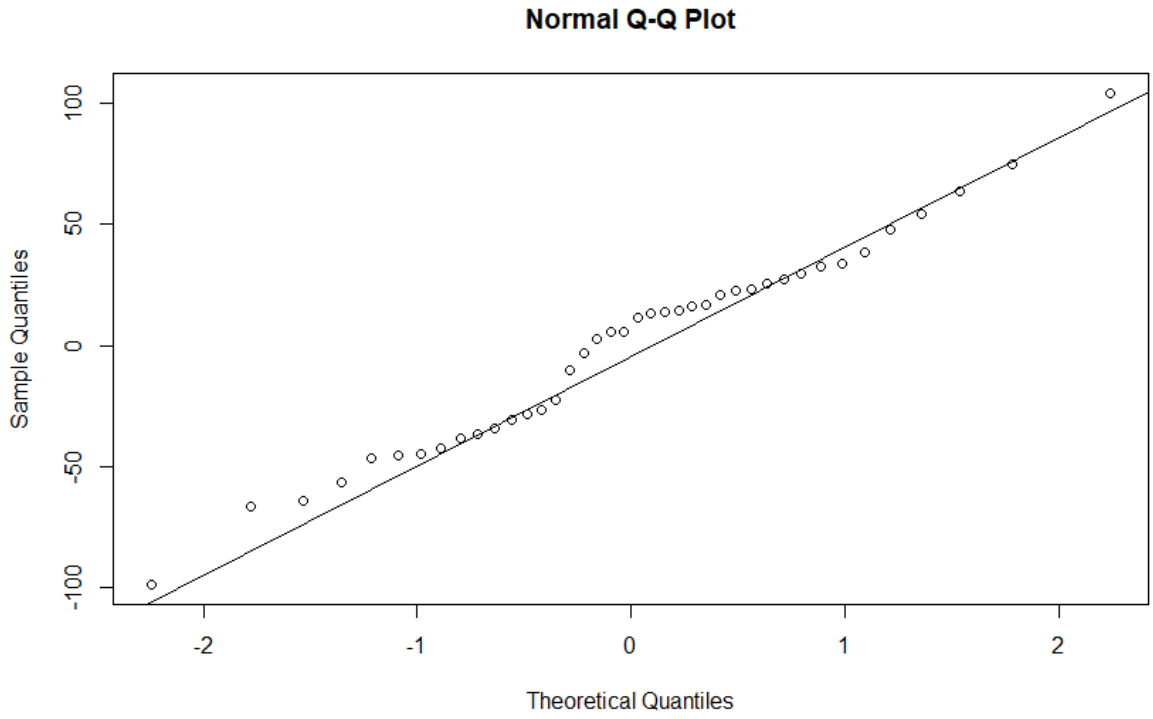
In both cases the lines are not parallel, indicating interaction.

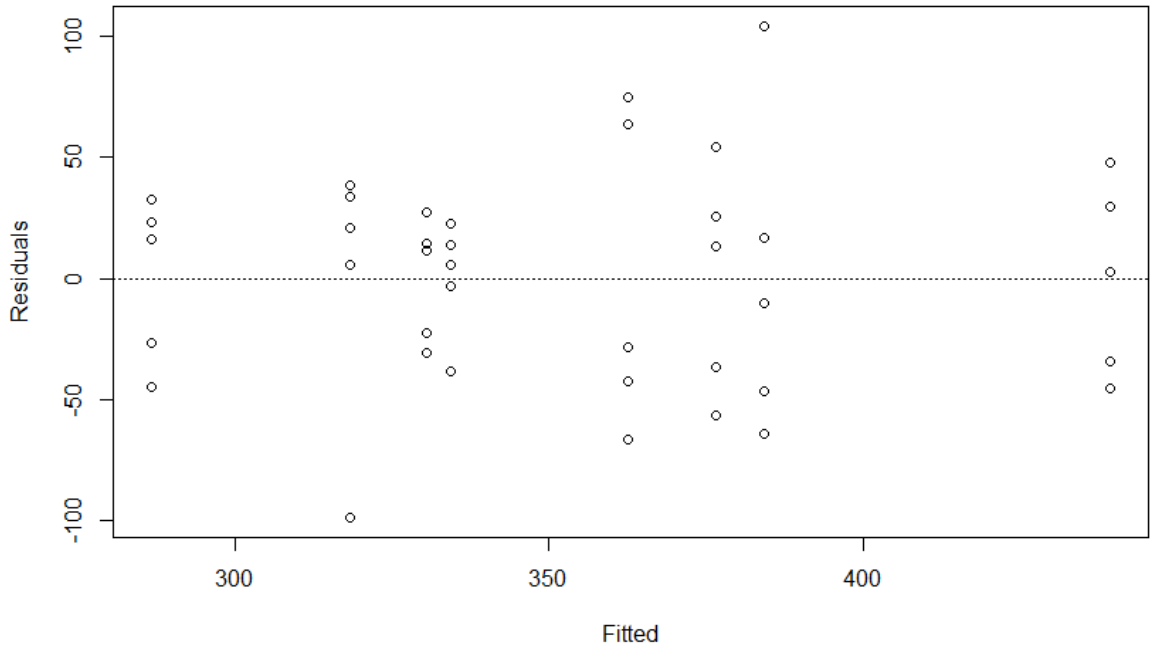
(c)

```
Call:
lm(formula = yield ~ variety0, data = oat)

Residuals:
   Min     1Q Median     3Q    Max
-98.40 -34.95   8.50  25.90 103.80

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   334.40      21.04  15.894  < 2e-16 ***
variety02      42.20      29.75   1.418  0.16578
variety03      28.20      29.75   0.948  0.35036
variety04     -47.60      29.75  -1.600  0.11949
variety05     105.00      29.75   3.529  0.00129 **
variety06      -3.80      29.75  -0.128  0.89918
variety07     -16.00      29.75  -0.538  0.59449
variety08      49.80      29.75   1.674  0.10394
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 47.05 on 32 degrees of freedom
Multiple R-squared:  0.5226,    Adjusted R-squared:  0.4181
F-statistic: 5.004 on 7 and 32 DF,  p-value: 0.0006568
```

Use variety1 as the baseline, and we can see there is enough evidence to conclude that there is difference between variety1 and variety5. Therefore, the yield of oats is affected by different varieties at 5% significance level.

(d)

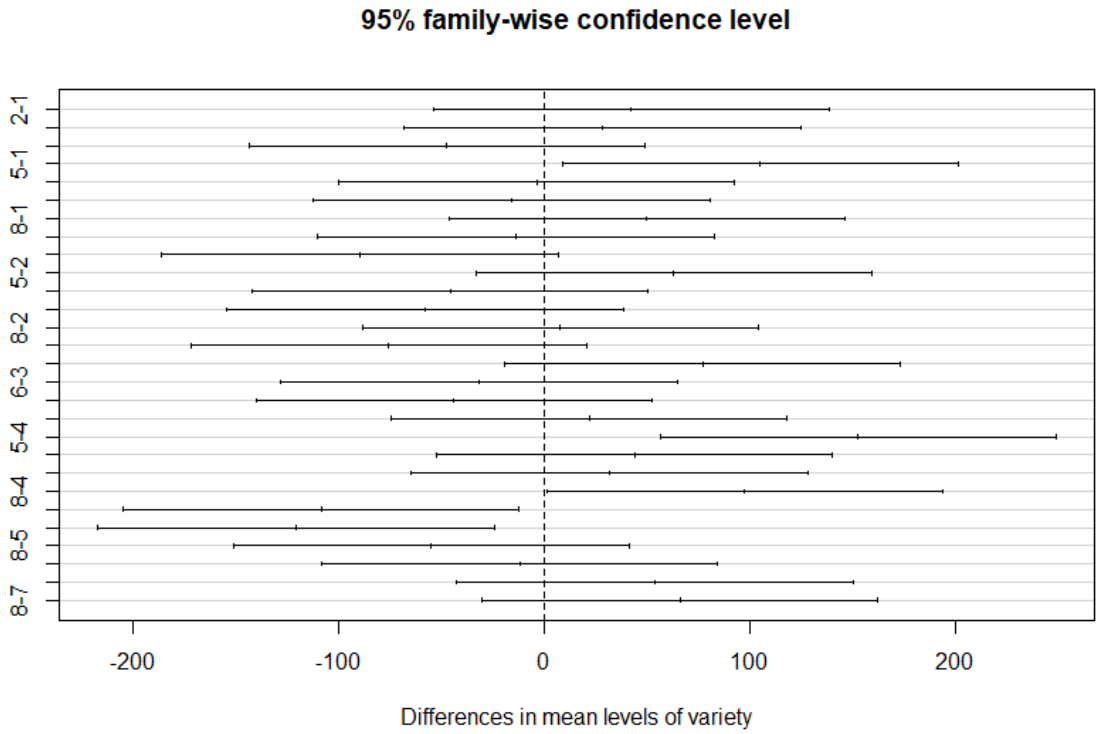




No, there isn't any unusual finding.

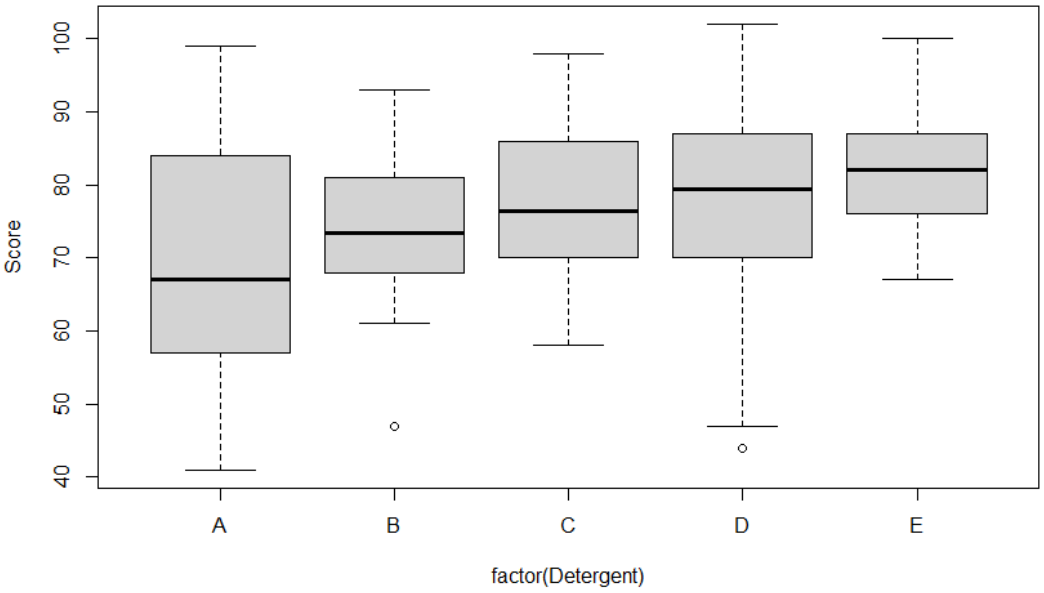
(e)

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = yield ~ variety)

$variety
      diff         lwr        upr     p adj
2-1   42.2  -54.185323 138.585323 0.8421823
3-1   28.2  -68.185323 124.585323 0.9784914
4-1  -47.6 -143.985323  48.785323 0.7469607
5-1  105.0    8.614677 201.385323 0.0249702
6-1   -3.8 -100.185323  92.585323 1.0000000
7-1  -16.0 -112.385323  80.385323 0.9993273
8-1   49.8  -46.585323 146.185323 0.7032342
3-2  -14.0 -110.385323  82.385323 0.9997200
4-2  -89.8 -186.185323   6.585323 0.0823960
5-2   62.8  -33.585323 159.185323 0.4299355
6-2  -46.0 -142.385323  50.385323 0.7771877
7-2  -58.2 -154.585323  38.185323 0.5251817
8-2    7.6  -88.785323 103.985323 0.9999955
4-3  -75.8 -172.185323  20.585323 0.2126443
5-3   76.8  -19.585323 173.185323 0.1999024
6-3  -32.0 -128.385323  64.385323 0.9574245
7-3  -44.2 -140.585323  52.185323 0.8092707
8-3   21.6  -74.785323 117.985323 0.9955018
5-4  152.6   56.214677 248.985323 0.0003306
6-4   43.8  -52.585323 140.185323 0.8160943
7-4   31.6  -64.785323 127.985323 0.9601392
8-4   97.4    1.014677 193.785323 0.0461777
6-5 -108.8 -205.185323 -12.414677 0.0181542
7-5 -121.0 -217.385323 -24.614677 0.0062594
8-5  -55.2 -151.585323  41.185323 0.5894148
7-6  -12.2 -108.585323  84.185323 0.9998881
8-6   53.6  -42.785323 149.985323 0.6236890
8-7   65.8  -30.585323 162.185323 0.3719006
```

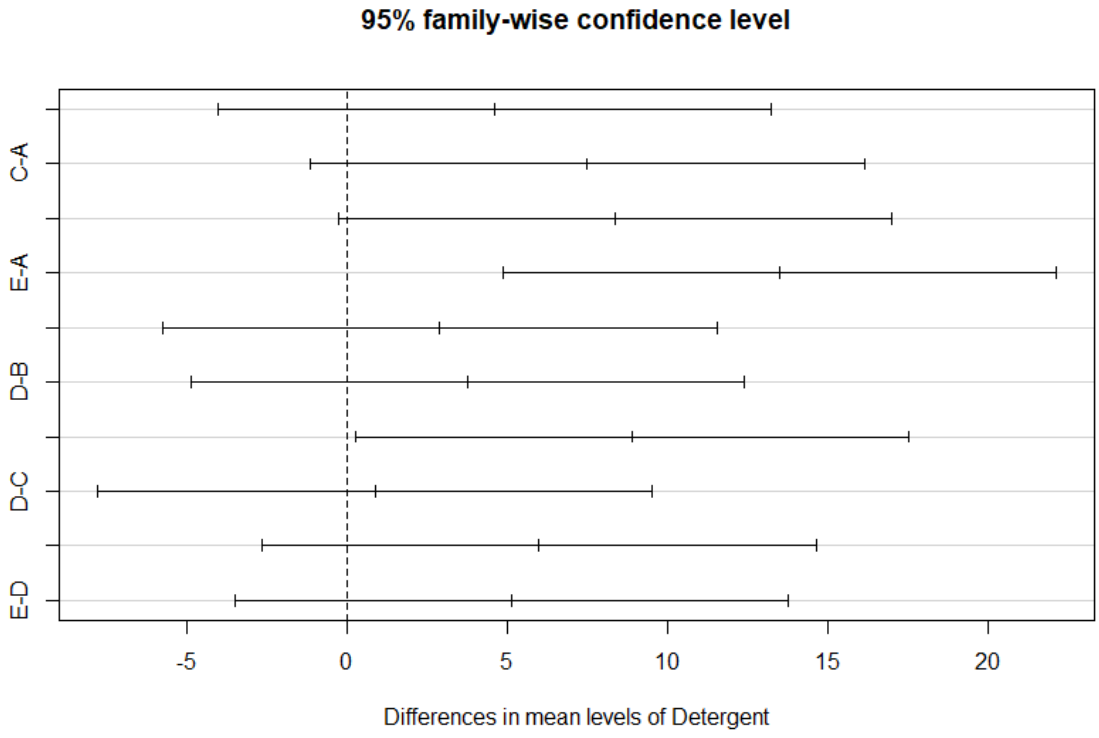


Yes, it's necessary. There are differences between (variety1 and variety5), (variety4 and variety5), (variety4 and variety7), (variety4 and variety8), (variety5 and variety6), (variety5 and variety7).

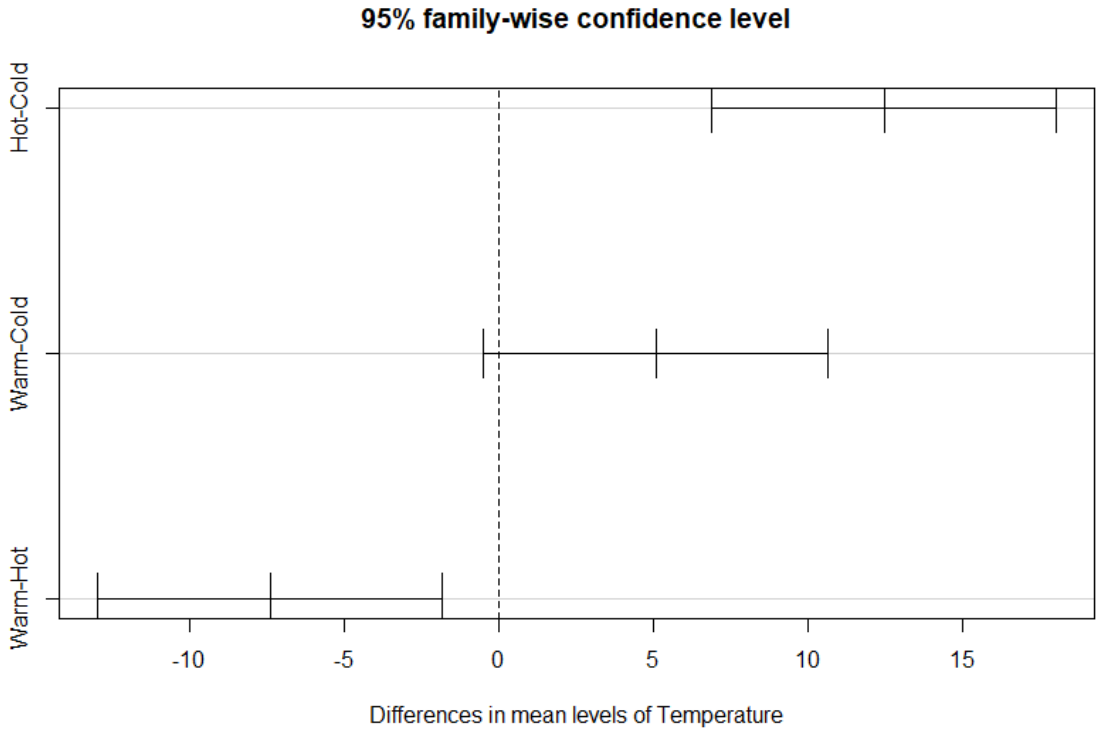
2.



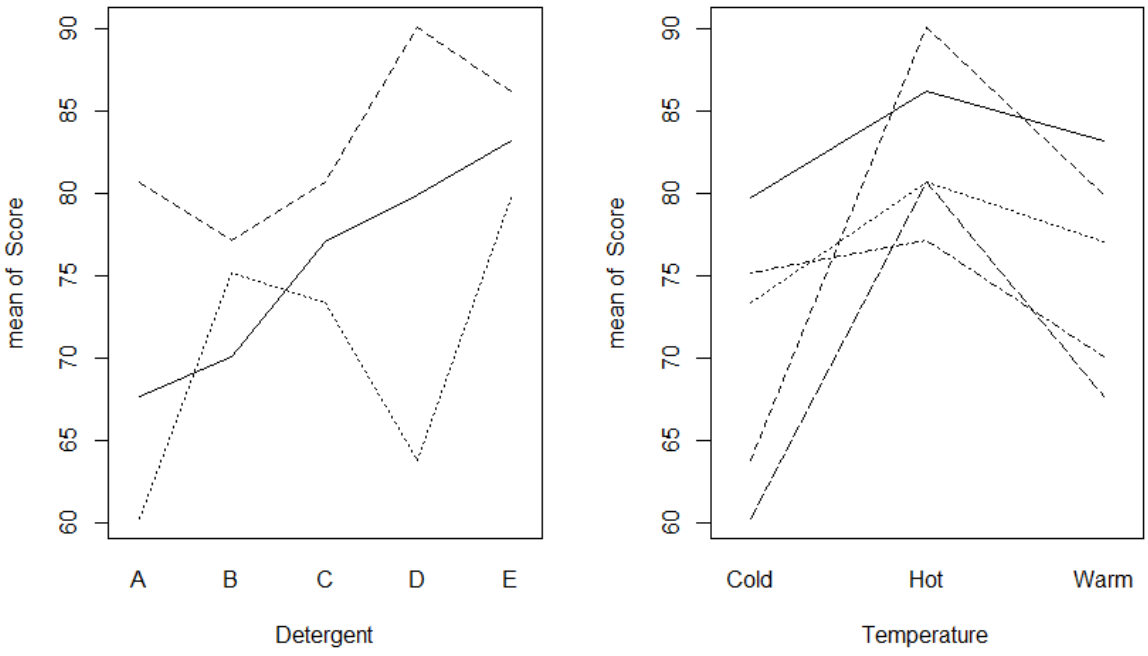
boxplot



There is sufficient statistical evidence to infer that there are differences in whiteness scores between (A and E) and (B and E) at 5% confidence level. However, There is no sufficient statistical evidence to infer that there are differences in whiteness scores between other detergents.



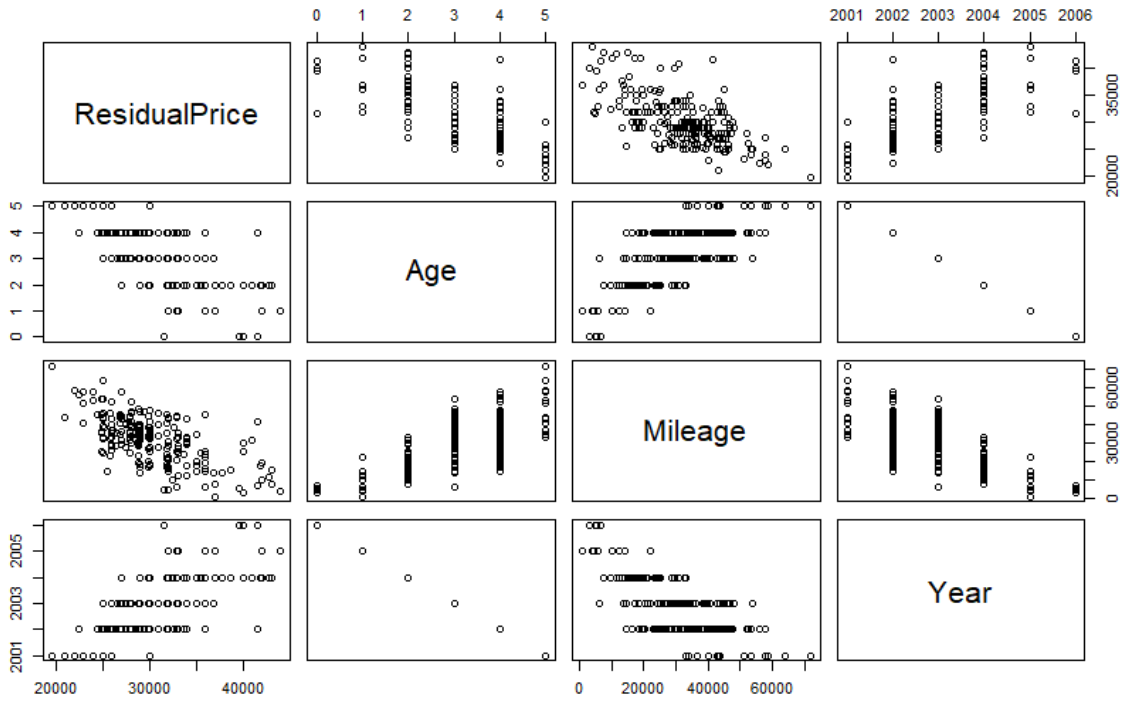
There is sufficient statistical evidence to infer that there are differences in whiteness scores between hot water and cold water.



In both cases the lines are not parallel, indicating interaction.

3.

(a)

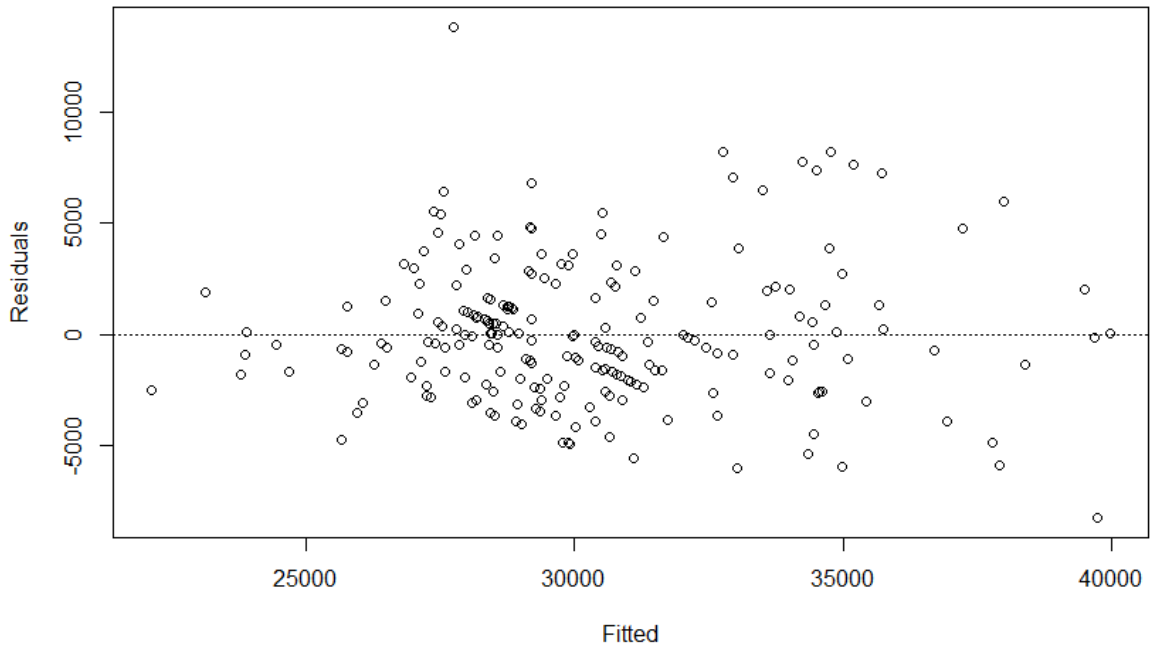


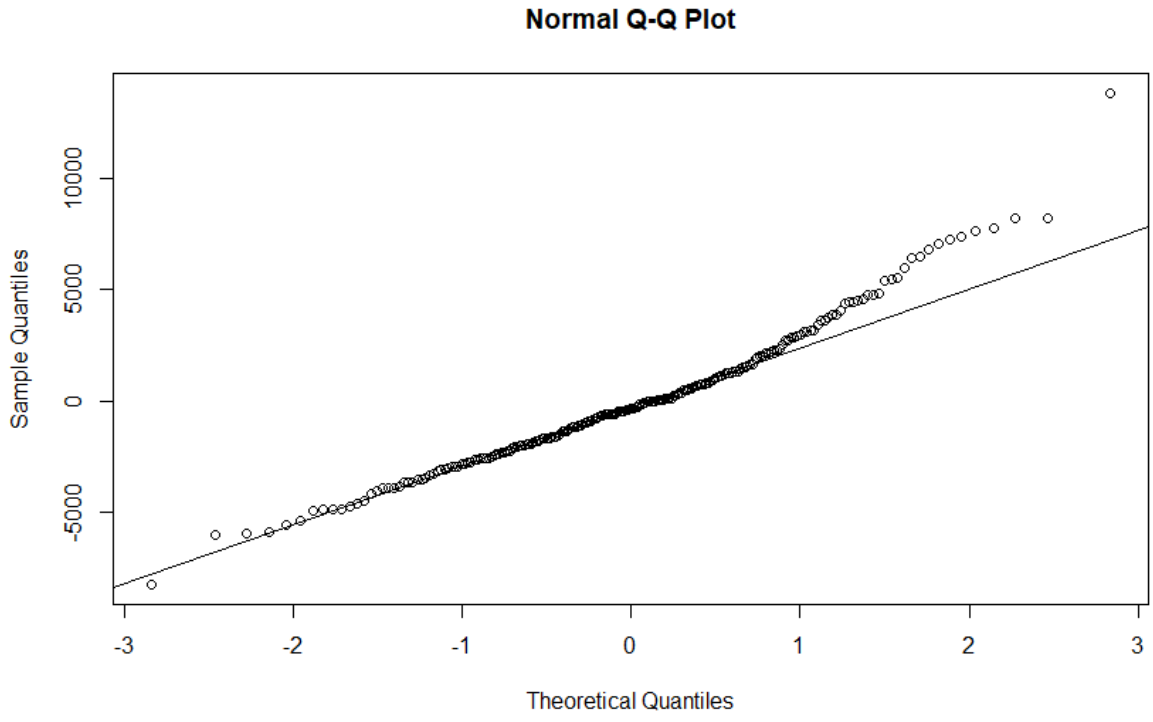
Yes, the relationships appear straight enough to permit using multiple linear regression with these variables. But the correlation between Age and Year is -1, so we need to drop one of them when we do multiple linear regression.

(b)

```
Call:
lm(formula = ResidualPrice ~ Age + Mileage, data = car)

Residuals:
    Min      1Q  Median      3Q     Max
-8245.5 -2052.5  -375.7  1525.4 13822.5

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.032e+04  7.218e+02  55.862  < 2e-16 ***
Age         -1.854e+03  2.889e+02  -6.417 8.72e-10 ***
Mileage     -1.240e-01  2.375e-02  -5.222 4.17e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3179 on 215 degrees of freedom
Multiple R-squared:  0.5104,    Adjusted R-squared:  0.5058
F-statistic: 112.1 on 2 and 215 DF,  p-value: < 2.2e-16
```

multiple linear regression model





From normal Q-Q plot, we can conclude this model doesn't meet the assumptions for multiple linear models.

(c)

```
                    2.5 %        97.5 %
(Intercept) 38901.1326609  4.174674e+04
Age         -2423.2011508 -1.284405e+03
Mileage        -0.1708354 -7.720998e-02
```

(d)

```
Call:
lm(formula = ResidualPrice ~ Age + Mileage, data = car)

Residuals:
    Min      1Q  Median      3Q     Max
-8245.5 -2052.5  -375.7  1525.4 13822.5

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.032e+04  7.218e+02  55.862  < 2e-16 ***
Age         -1.854e+03  2.889e+02  -6.417 8.72e-10 ***
Mileage     -1.240e-01  2.375e-02  -5.222 4.17e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3179 on 215 degrees of freedom
Multiple R-squared:  0.5104,    Adjusted R-squared:  0.5058
F-statistic: 112.1 on 2 and 215 DF,  p-value: < 2.2e-16
```

The R-squared value:0.5104 means that there are about 51% of the variations can be explained by this model.

The p-value of Age is smaller than 0.05, which means we can reject the null hypothesis: coefficient of Age = 0. When Age increases one, the predicted price would decrease 1.854e+03.

The p-value of Mileage is smaller than 0.05, which means we can reject the null hypothesis: coefficient of Mileage = 0. When Mileage increases one, the predicted price would decrease 1.240e-01.

(e)

1. The brand and the quality of car
2. The habits of former driver
3. Whether there is an accident on it