STAT2170 - Assignment 1

Theodore Harding - 45234671

Due Date: 2023-10-27

Question 1

Load the traffic data

```
traffic <- read.csv("traffic.csv")
```

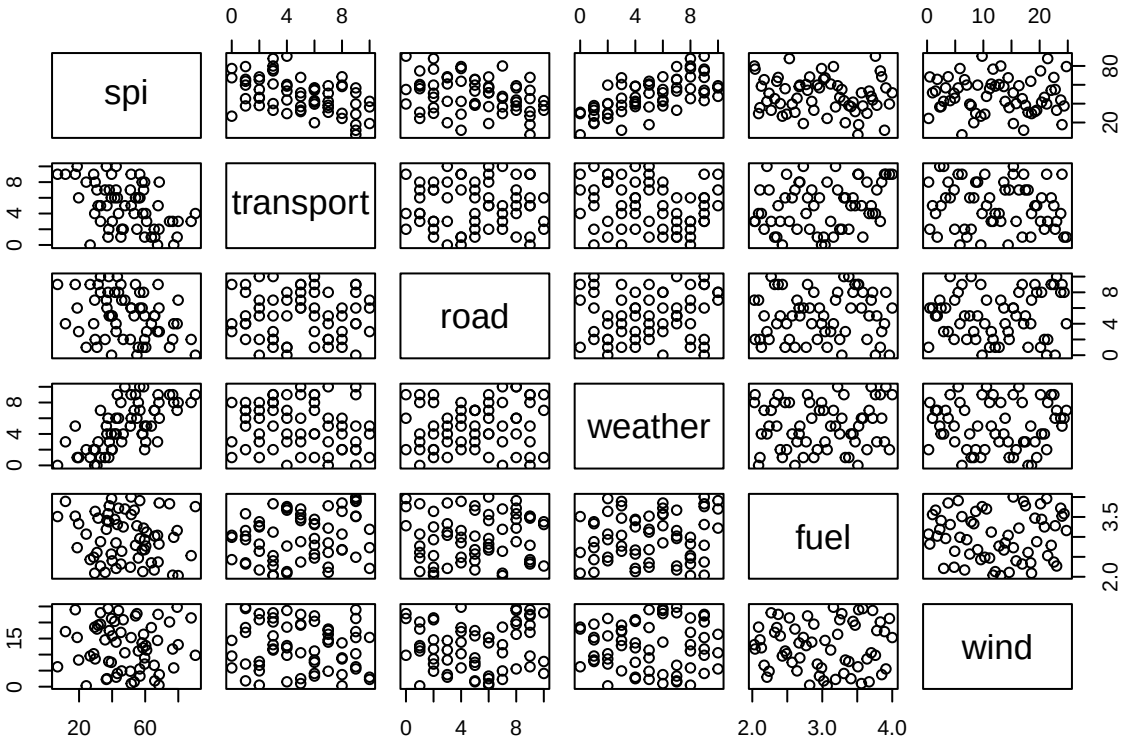
Part A

Calculate correlation matrix and create a scatterplot matrix

```
cor(traffic)
```

```
##                   spi    transport         road     weather         fuel
## spi        1.00000000 -0.472909967 -0.303836850  0.66672345 -0.138153417
## transport -0.47290997  1.000000000 -0.005714728 -0.16971072  0.240947972
## road      -0.30383685 -0.005714728  1.000000000  0.12495993  0.043675635
## weather    0.66672345 -0.169710717  0.124959926  1.00000000  0.110531767
## fuel      -0.13815342  0.240947972  0.043675635  0.11053177  1.000000000
## wind      -0.03466263 -0.131014749  0.080481857  0.00751783  0.006532832
##                   wind
## spi       -0.034662632
## transport -0.131014749
## road       0.080481857
## weather    0.007517830
## fuel       0.006532832
## wind       1.000000000
```

```
pairs(traffic)
```



Part B

Fit a linear regression model

```
fit <- lm(spi ~ ., data = traffic)
summary(fit)
```

```
##
## Call:
## lm(formula = spi ~ ., data = traffic)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.1596  -4.9415   0.1278   5.1686  21.7415
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  62.8071     7.4080   8.478 1.27e-11 ***
## transport    -2.1750     0.4611  -4.717 1.63e-05 ***
## road         -2.4097     0.4365  -5.520 9.04e-07 ***
## weather       4.2456     0.4473   9.492 2.92e-13 ***
## fuel         -3.6145     2.2759  -1.588    0.118
## wind         -0.1358     0.1764  -0.769    0.445
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.913 on 56 degrees of freedom
## Multiple R-squared:  0.7405, Adjusted R-squared:  0.7174
## F-statistic: 31.96 on 5 and 56 DF,  p-value: 3.039e-15
```

Calculate confidence interval for the 'weather' variable

```
confint(fit, 'weather', level = 0.95)
```

```
##            2.5 %   97.5 %
## weather 3.349648 5.141639
```

Part C

Display the summary of the linear regression model

```
summary(fit)
```

```
##
## Call:
## lm(formula = spi ~ ., data = traffic)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.1596  -4.9415   0.1278   5.1686  21.7415
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  62.8071     7.4080   8.478 1.27e-11 ***
## transport    -2.1750     0.4611  -4.717 1.63e-05 ***
## road         -2.4097     0.4365  -5.520 9.04e-07 ***
## weather       4.2456     0.4473   9.492 2.92e-13 ***
## fuel         -3.6145     2.2759  -1.588    0.118
## wind         -0.1358     0.1764  -0.769    0.445
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.913 on 56 degrees of freedom
## Multiple R-squared:  0.7405, Adjusted R-squared:  0.7174
## F-statistic: 31.96 on 5 and 56 DF,  p-value: 3.039e-15
```

Present the model equation

$spi = 62.8071 - 2.1750 * transport - 2.4097 * road + 4.2456 * weather - 3.6145 * fuel - 0.1358 * wind$

Null hypothesis: $Beta1 = Beta2 = ... = BetaN = 0$ Alternate hypothesis: $Atleast1Beta(i)! = 0$

Anova table for model comparison

```r
traffic$null <- mean(traffic$spi)
null_fit <- lm(spi ~ null, data = traffic)
summary(null_fit)
```

```
##
## Call:
## lm(formula = spi ~ null, data = traffic)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -41.694 -12.086  -1.479  11.471  41.656
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   48.694      2.368   20.56   <2e-16 ***
## null              NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.65 on 61 degrees of freedom
```

```r
anova(null_fit, fit)
```

```
## Analysis of Variance Table
##
## Model 1: spi ~ null
## Model 2: spi ~ transport + road + weather + fuel + wind
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1     61 21206.2
## 2     56  5502.6  5     15704 31.963 3.039e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Part D

Goodness of fit

```r
summary(fit)
```

```
##
## Call:
## lm(formula = spi ~ ., data = traffic)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.1596  -4.9415   0.1278   5.1686  21.7415
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  62.8071     7.4080   8.478 1.27e-11 ***
```

```
## transport    -2.1750     0.4611  -4.717 1.63e-05 ***
## road         -2.4097     0.4365  -5.520 9.04e-07 ***
## weather       4.2456     0.4473   9.492 2.92e-13 ***
## fuel         -3.6145     2.2759  -1.588    0.118
## wind         -0.1358     0.1764  -0.769    0.445
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.913 on 56 degrees of freedom
## Multiple R-squared:  0.7405, Adjusted R-squared:  0.7174
## F-statistic: 31.96 on 5 and 56 DF,  p-value: 3.039e-15
```

Part E

R-squared is a measure of the model fit and explained variability

```
summary(fit)
```

```
##
## Call:
## lm(formula = spi ~ ., data = traffic)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.1596  -4.9415   0.1278   5.1686  21.7415
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  62.8071     7.4080   8.478 1.27e-11 ***
## transport    -2.1750     0.4611  -4.717 1.63e-05 ***
## road         -2.4097     0.4365  -5.520 9.04e-07 ***
## weather       4.2456     0.4473   9.492 2.92e-13 ***
## fuel         -3.6145     2.2759  -1.588    0.118
## wind         -0.1358     0.1764  -0.769    0.445
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.913 on 56 degrees of freedom
## Multiple R-squared:  0.7405, Adjusted R-squared:  0.7174
## F-statistic: 31.96 on 5 and 56 DF,  p-value: 3.039e-15
```

Part F (Stepwise Selection)

Forward stepwise selection

```
fit <- lm(spi ~ transport + road + weather + fuel, data = traffic)
summary(fit)
```

```
##
## Call:
## lm(formula = spi ~ transport + road + weather + fuel, data = traffic)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.9347  -4.2440   0.0528   5.0544  21.4515
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  61.1610     7.0669   8.655 5.69e-12 ***
## transport    -2.1257     0.4550  -4.672 1.86e-05 ***
## road         -2.4372     0.4335  -5.622 5.92e-07 ***
## weather       4.2565     0.4454   9.555 1.94e-13 ***
## fuel         -3.6853     2.2659  -1.626    0.109
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.877 on 57 degrees of freedom
## Multiple R-squared:  0.7378, Adjusted R-squared:  0.7194
## F-statistic: 40.09 on 4 and 57 DF,  p-value: 5.959e-16
```

Backward stepwise selection (removing 'fuel')

```
fit <- lm(spi ~ transport + road + weather, data = traffic)
summary(fit)
```

```
##
## Call:
## lm(formula = spi ~ transport + road + weather, data = traffic)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.672  -5.643   1.067   4.656  23.164
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  51.7370     4.1027  12.611  < 2e-16 ***
## transport    -2.3216     0.4449  -5.218 2.54e-06 ***
## road         -2.4563     0.4394  -5.590 6.40e-07 ***
## weather       4.1450     0.4463   9.286 4.48e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.02 on 58 degrees of freedom
## Multiple R-squared:  0.7256, Adjusted R-squared:  0.7114
## F-statistic: 51.12 on 3 and 58 DF,  p-value: 2.724e-16
```

Part G

Adjusted R^2 takes into account the number of variables in the dataset

Question 2

```
cake <- read.csv("cake.csv")
```

Part A

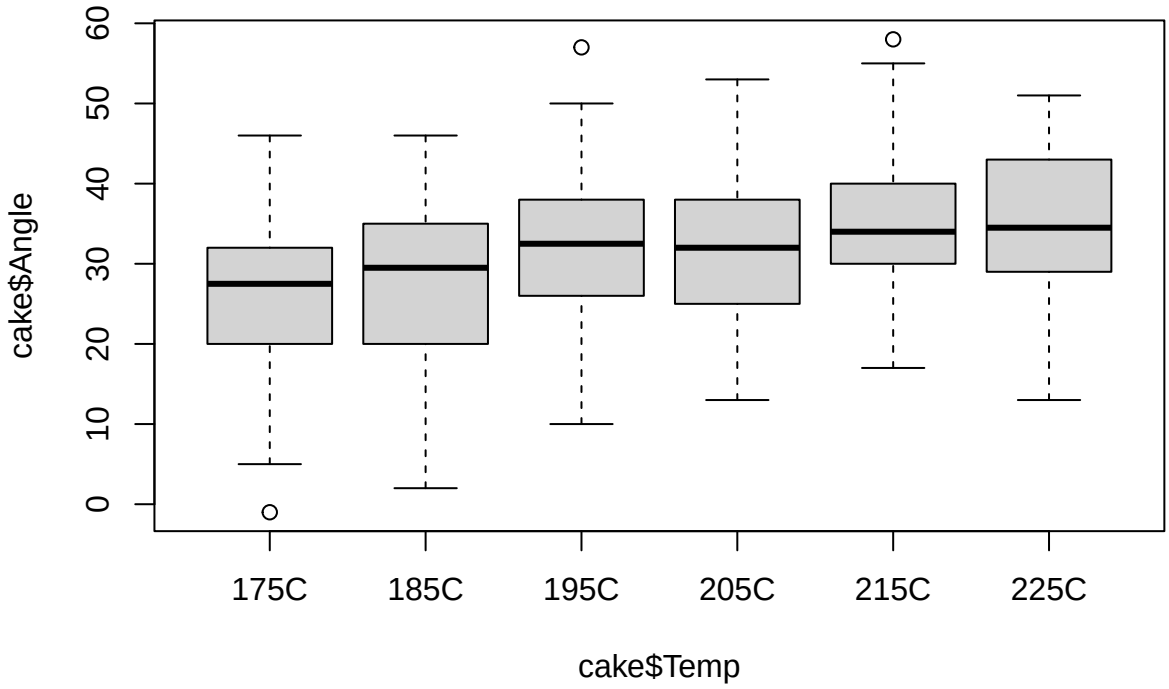
```
table(cake$Temp, cake$Recipe)
```

```
##
##         A  B  C
##   175C 14 14 14
##   185C 14 14 14
##   195C 14 14 14
##   205C 14 14 14
##   215C 14 14 14
##   225C 14 14 14
```

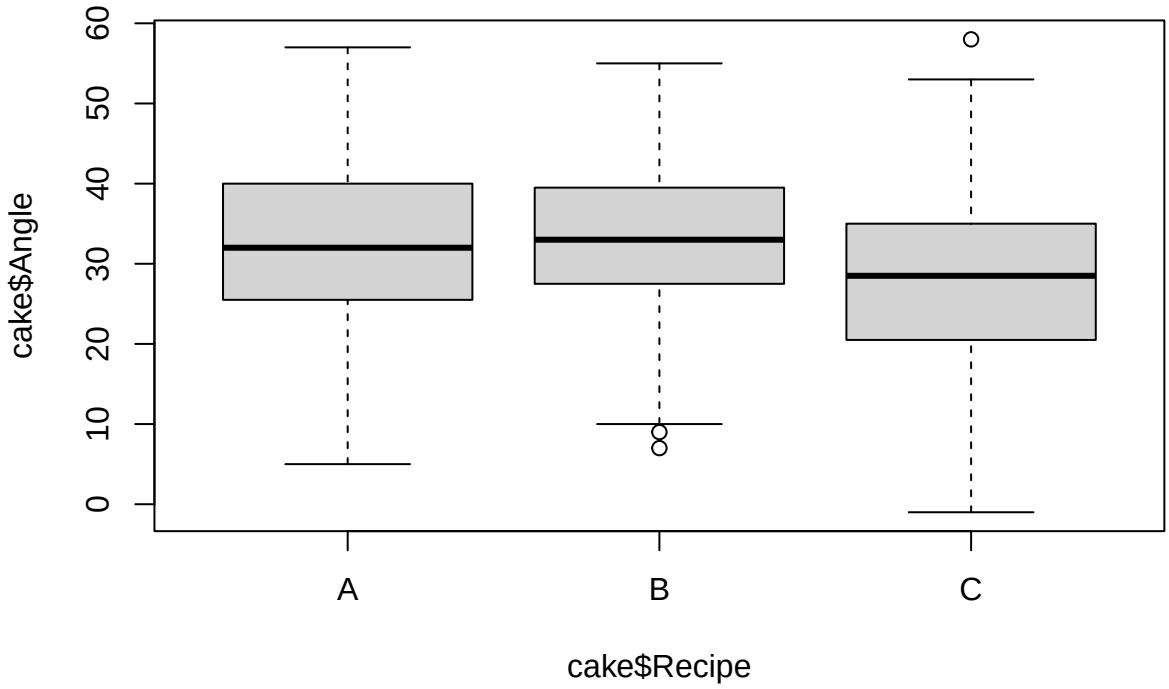
Temp and recipe do have an impact on the angle the cake broke but the interaction variable is insignificant It is balanced - there is the same number of subjects in each cohort

Part B

```
boxplot(cake$Angle~cake$Temp)
```




```
boxplot(cake$Angle~cake$Recipe)
```



Part C

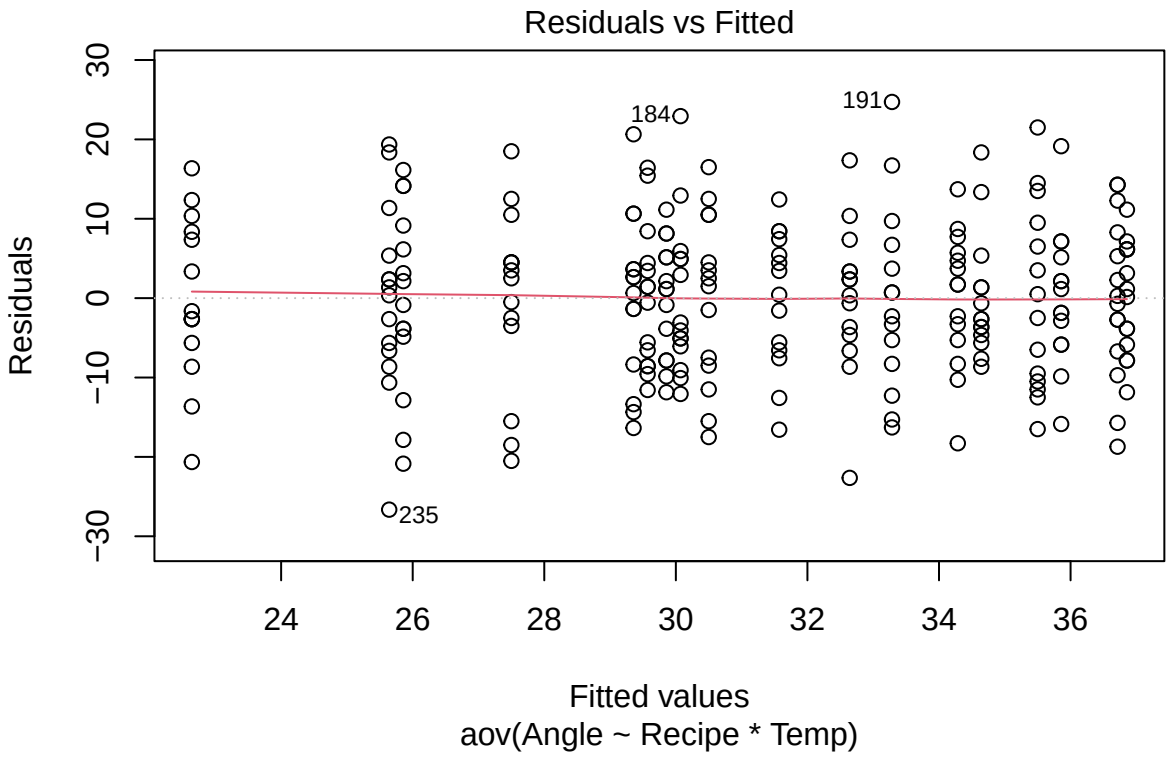
```
av = aov(formula = Angle~Recipe*Temp, data = cake)
```

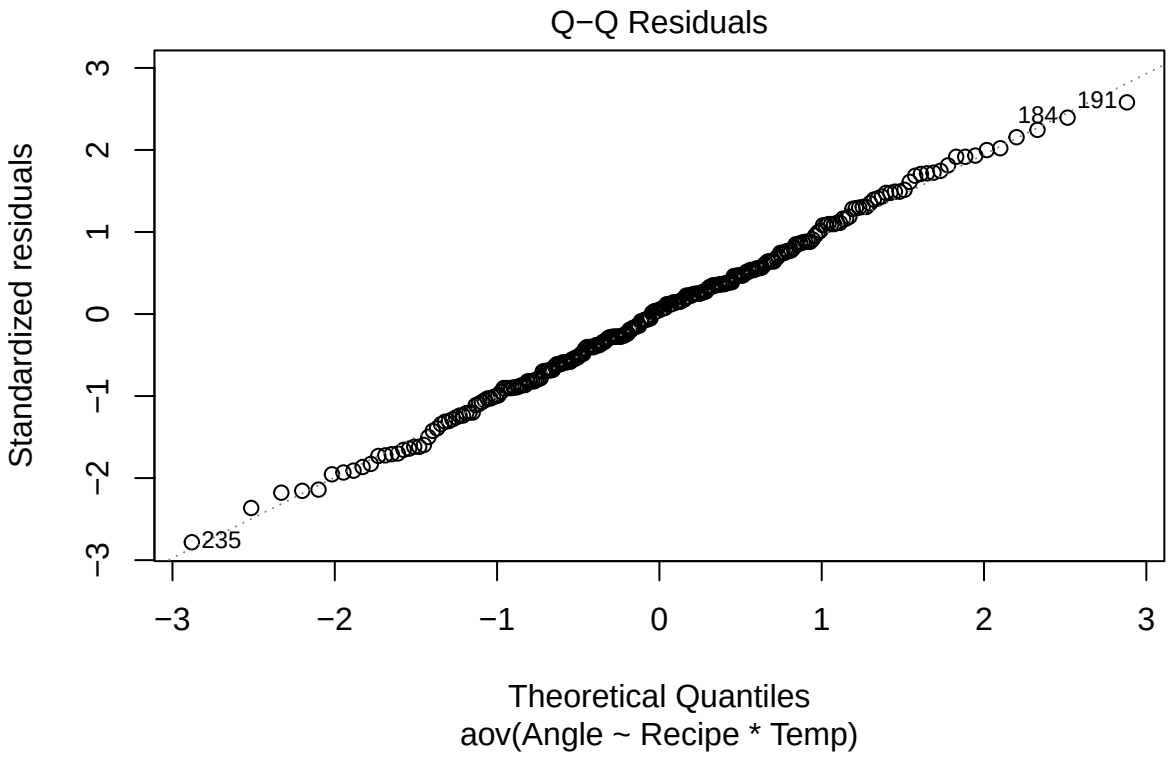
Part D

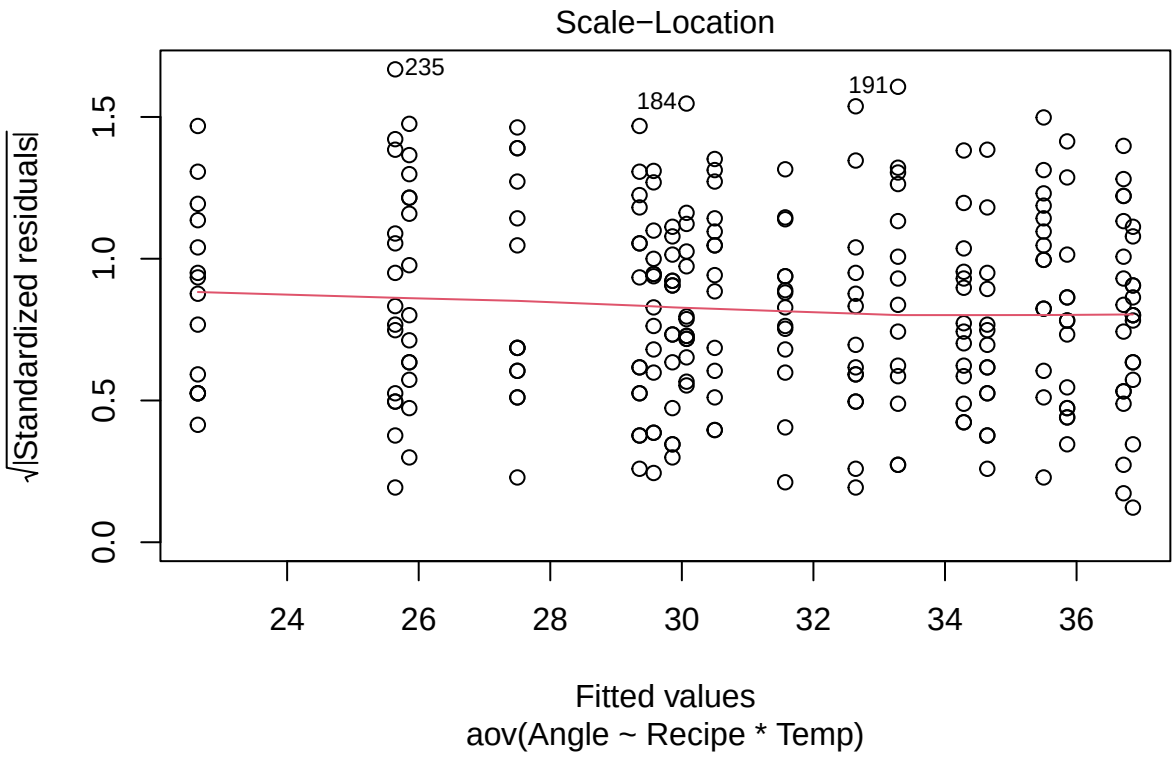
```
summary(av)
```

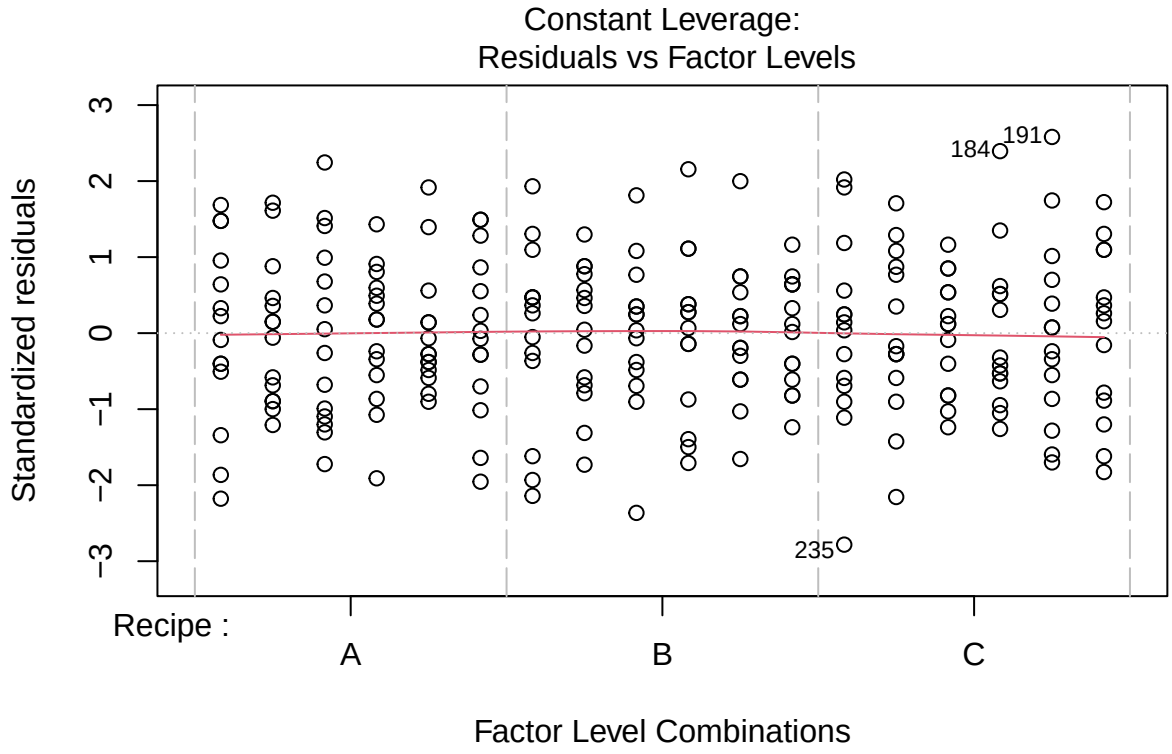
```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## Recipe        2    845   422.4   4.276 0.014998 *
## Temp          5   2530   506.0   5.123 0.000177 ***
## Recipe:Temp  10    636    63.6   0.643 0.775632
## Residuals   234  23114    98.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
plot(av)
```


Residuals vs Fitted
Residuals
30
20
10
0
−10
−30
184
191
235
24
26
28
30
32
34
36
Fitted values
aov(Angle ~ Recipe * Temp)


Q−Q Residuals
Standardized residuals
Theoretical Quantiles
aov(Angle ~ Recipe * Temp)
235
184
191
−3
−2
−1
0
1
2
3


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
235
184
191
24
26
28
30
32
34
36
Fitted values
aov(Angle ~ Recipe * Temp)

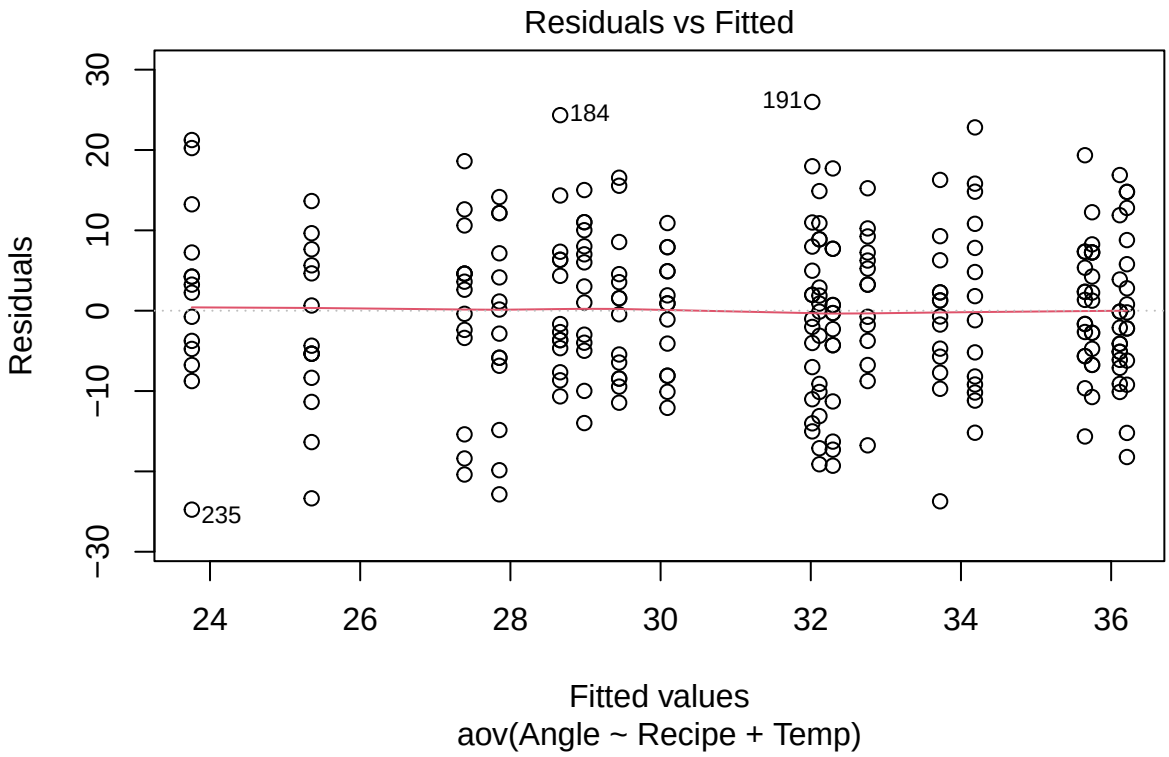


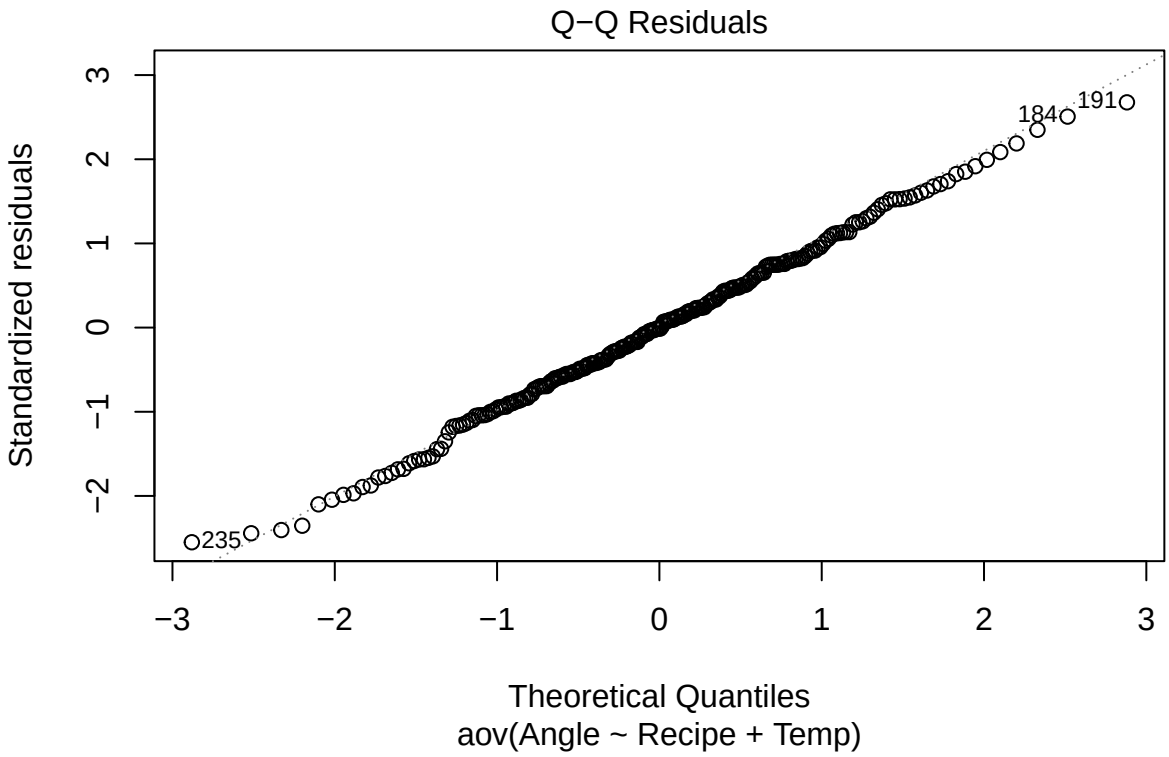
Part E

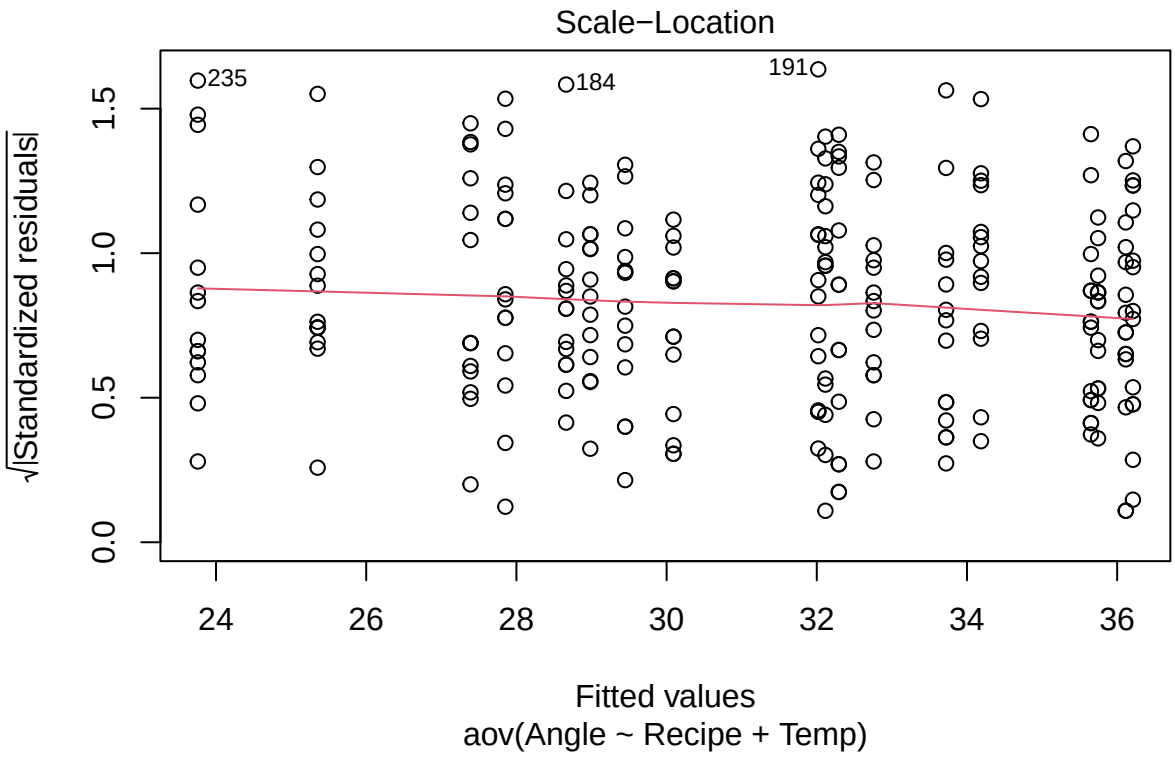
```
me <- aov(formula = Angle~Recipe+Temp, data = cake)
summary(me)
```

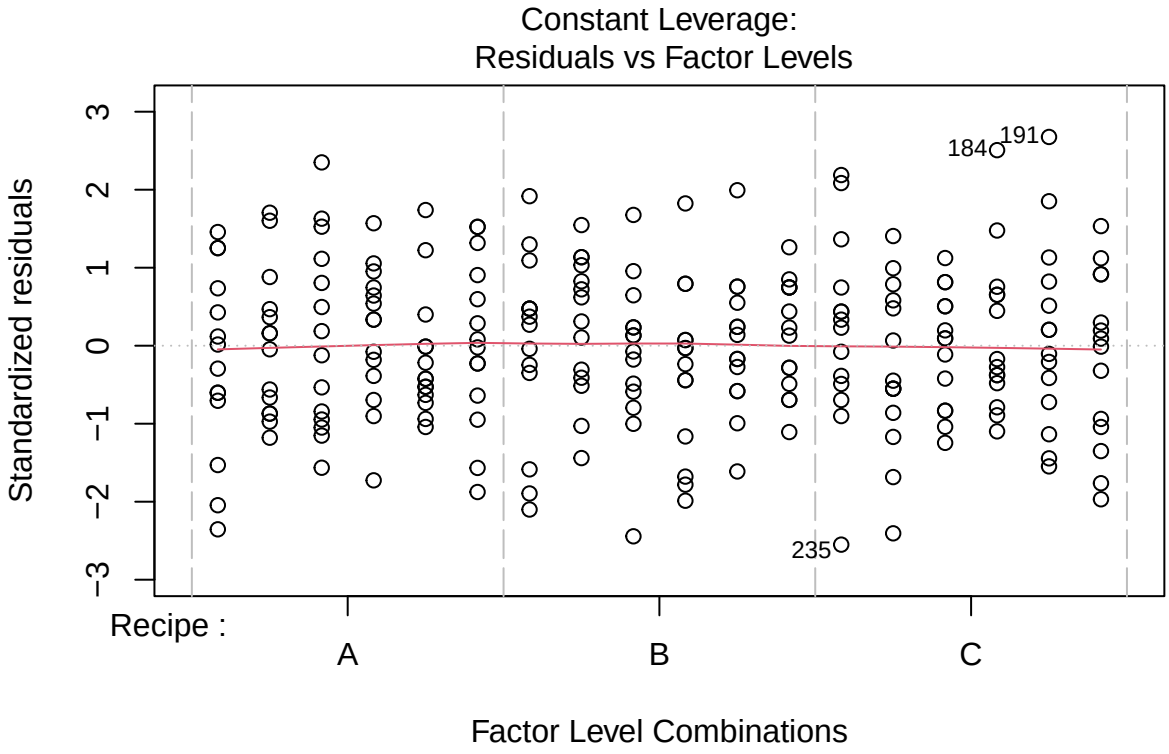
```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## Recipe        2    845   422.4   4.340 0.014064 *
## Temp          5   2530   506.0   5.199 0.000149 ***
## Residuals   244  23749    97.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
plot(me)
```


Residuals vs Fitted
Residuals
30
20
10
0
−10
−30
184
191
235
24
26
28
30
32
34
36
Fitted values
aov(Angle ~ Recipe + Temp)


Q−Q Residuals
Standardized residuals
Theoretical Quantiles
aov(Angle ~ Recipe + Temp)
235
184
191
−3
−2
−1
0
1
2
3


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
235
184
191
24
26
28
30
32
34
36
Fitted values
aov(Angle ~ Recipe + Temp)


Constant Leverage:
Residuals vs Factor Levels
Standardized residuals
3
2
1
0
−1
−2
−3
184
191
235
Recipe :
A
B
C
Factor Level Combinations