1 Tooth growth in guinea pigs

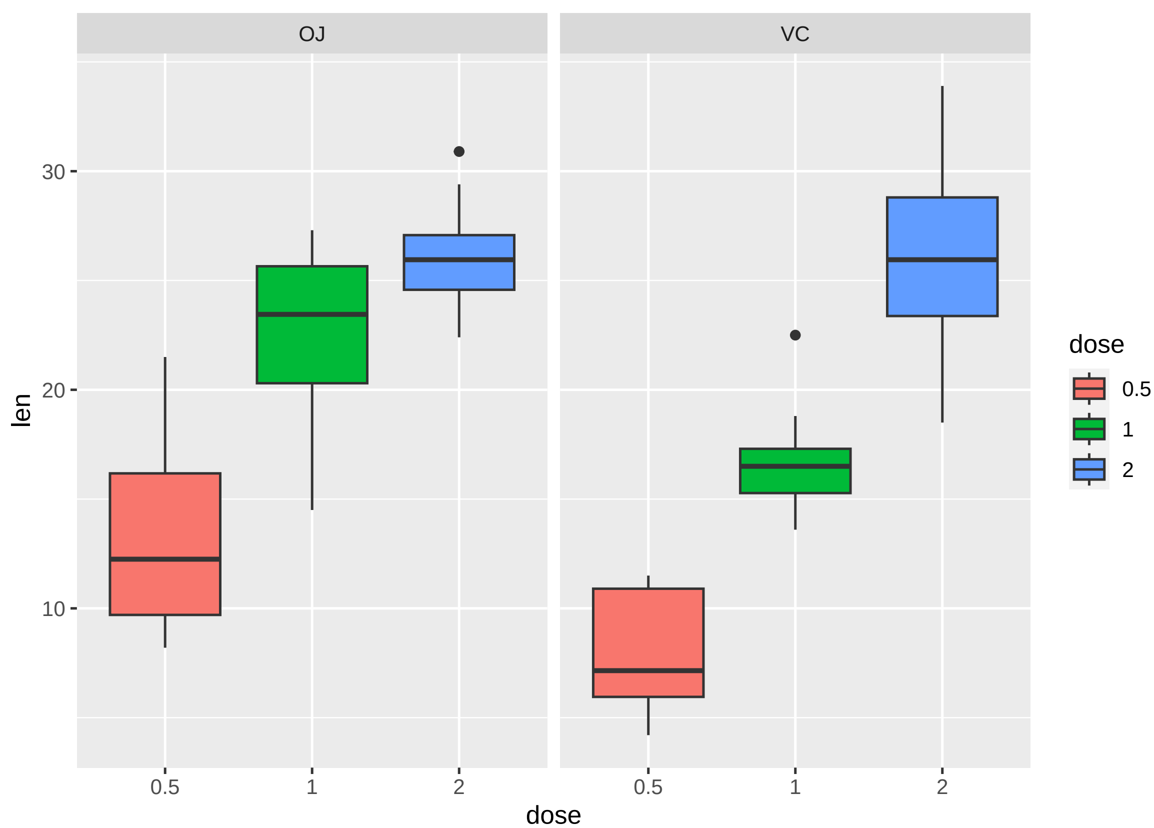
Q1.1: Download the *ToothGrowth* data set from R and read its description.

(a) Create some plots which would help you to explore the effects of vitamin C dose and its delivery method on the tooth growth and state an appropriate research hypothesis.

(b) Consider *dose* as a factor variable and use the analysis of variance to test your hypothesis. Are the dose effect and delivery effect significant?

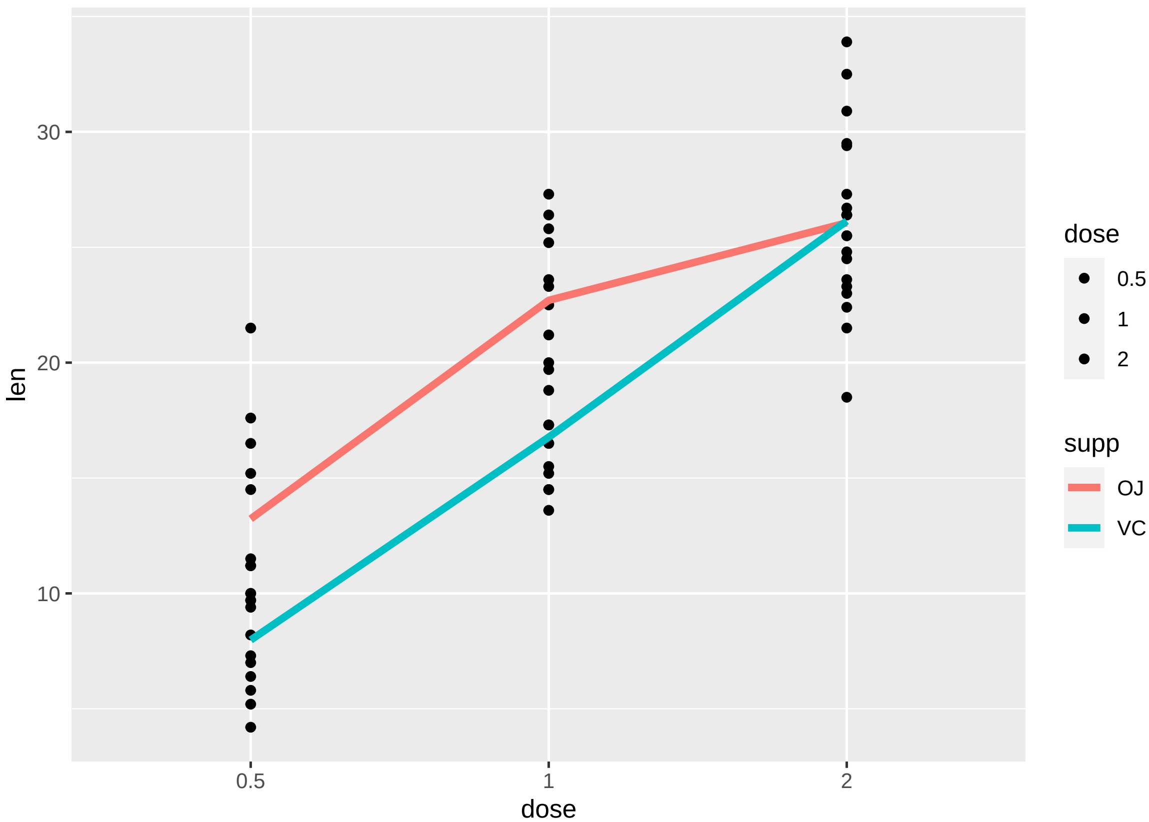
(c) Which statistical test can we use to check whether the dose effect varies depending on the delivery method?

```
data("ToothGrowth");
library(ggplot2)

## Warning:  Paket 'ggplot2' wurde unter R Version 4.2.2 erstellt

ToothGrowth$dose <- as.factor(ToothGrowth$dose);
gg.base <- ggplot(ToothGrowth,aes(x = dose, y = len, fill = dose )) +   geom_boxplot();
  gg.base + facet_wrap( ~ supp);
```




```
gg.base.1 <- ggplot(ToothGrowth,aes(x = dose, y = len, fill = dose )) +  geom_point();
gg.base.1 +  stat_summary(aes(group = supp, color = supp), geom = "line", fun.y = mean, size = 1.5);
```

```
## Warning:  The `fun.y` argument of `stat_summary()` is deprecated as of ggplot2 3.3.0.
## i Please use the `fun` argument instead.
## Warning:  Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
```




```
## The mean effects model:
mod1 <- lm(len ~ as.factor(dose) + supp, data=ToothGrowth);
anova(mod1)

## Analysis of Variance Table
##
## Response: len
##                 Df  Sum Sq Mean Sq F value    Pr(>F)
## as.factor(dose)  2 2426.43 1213.22  82.811 < 2.2e-16 ***
## supp             1  205.35  205.35  14.017 0.0004293 ***
## Residuals       56  820.43   14.65
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## The interaction model:
mod2 <- lm(len ~ as.factor(dose) * supp, data=ToothGrowth);
anova(mod2)

## Analysis of Variance Table
##
```

```
## Response: len
##                       Df  Sum Sq Mean Sq F value    Pr(>F)
## as.factor(dose)        2 2426.43 1213.22  92.000 < 2.2e-16 ***
## supp                   1  205.35  205.35  15.572 0.0002312 ***
## as.factor(dose):supp  2  108.32   54.16   4.107 0.0218603 *
## Residuals             54  712.11   13.19
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Model comparison, F-test of dose*supplement interaction:
anova(mod1,mod2)

## Analysis of Variance Table
##
## Model 1: len ~ as.factor(dose) + supp
## Model 2: len ~ as.factor(dose) * supp
##   Res.Df    RSS Df Sum of Sq     F  Pr(>F)
## 1     56 820.43
## 2     54 712.11  2    108.32 4.107 0.02186 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2 Weight change on different diets

Q2.1: The data set is taken from the website: `https://rcompanion.org/handbook/G_09.html`. In order to conduct a (hypothetical) study about the association between eating habits and weight change, a (hypothetical) researcher enrolled 27 randomly selected participants (9 from each USA, UK, New Zealand), and using the blocks of size = 3, randomly assigned them to the experimental diet A, B, C. The resulting data set is given in the file `DietWeight.txt`.

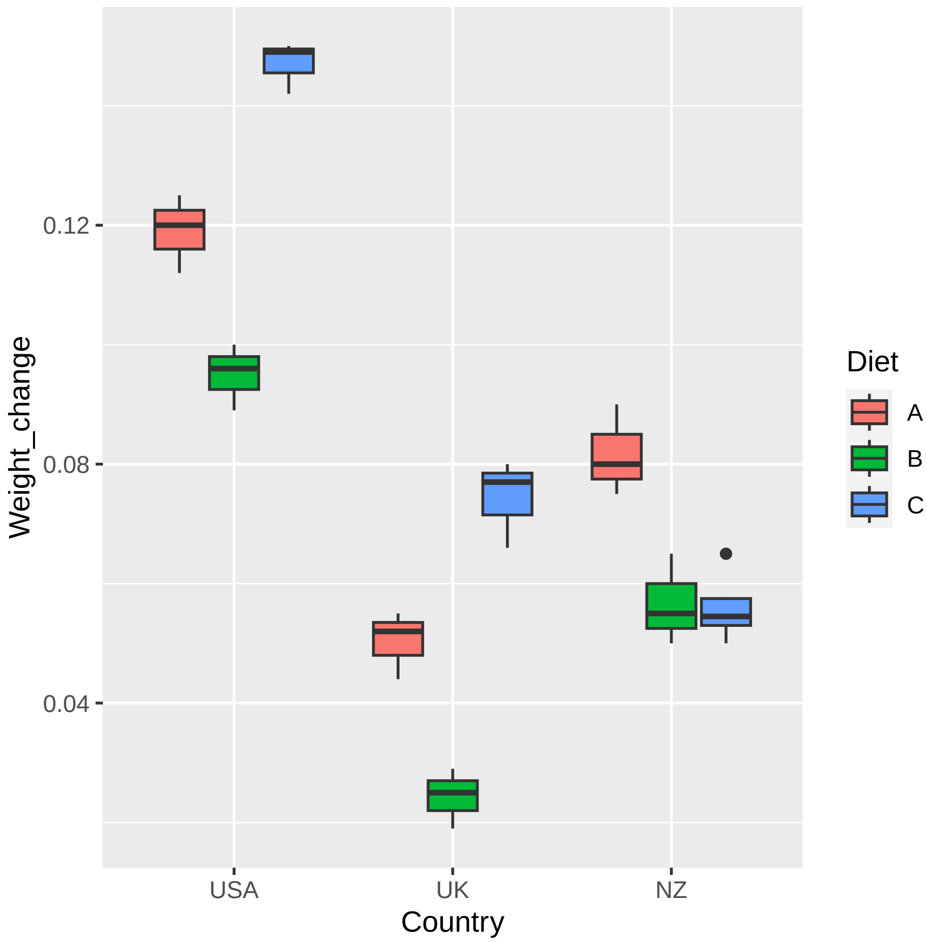
(a) Download the data and make some plots which can help to explore the interaction between Country and Diet.

(b) Do the plots suggest that the effect of diet is not consistent across all three countries?

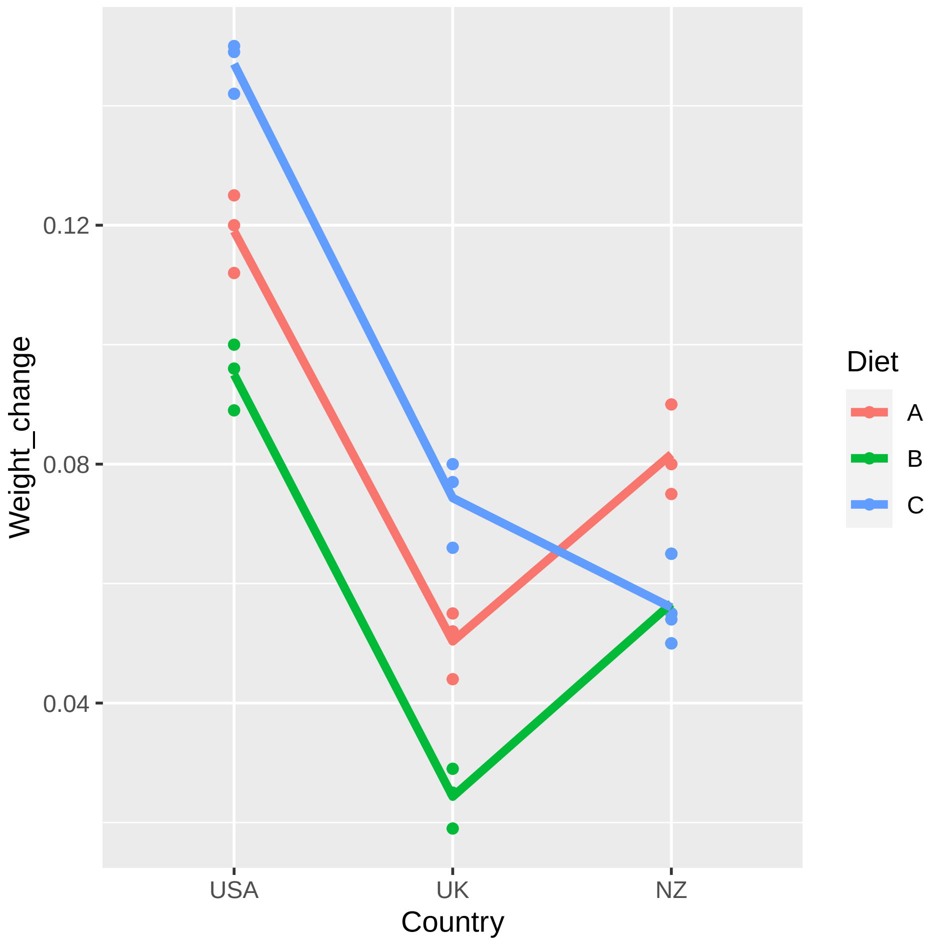
(c) Carry out a formal statistical test to get evidence supporting your conclusion. Were the country wise differences significant?

```
Input <- ("
Diet    Country  Weight_change
 A       USA      0.120
 A       USA      0.125
 A       USA      0.112
 A       UK       0.052
 A       UK       0.055
 A       UK       0.044
 A       NZ       0.080
 A       NZ       0.090
 A       NZ       0.075
 B       USA      0.096
```

```
 B       USA      0.100
 B       USA      0.089
 B       UK       0.025
 B       UK       0.029
 B       UK       0.019
 B       NZ       0.055
 B       NZ       0.065
 B       NZ       0.050
 C       USA      0.149
 C       USA      0.150
 C       USA      0.142
 C       UK       0.077
 C       UK       0.080
 C       UK       0.066
 C       NZ       0.055
 C       NZ       0.065
 C       NZ       0.050
 C       NZ       0.054
")

## to read the data directly:
my.data <- read.table(textConnection(Input), header=TRUE)
## to read the data from a file:
my.data <- read.table("data/DietWeight.txt", sep = ",", header=TRUE)

### Order levels of the factor; otherwise R will alphabetize them
my.data$Country <- factor(my.data$Country,
                      levels=unique(my.data$Country))

library(ggplot2)

gg.base <- ggplot(my.data,aes(x = Country, y = Weight_change, fill = Diet )) +   geom_boxplot();
gg.base;
```




```
gg.base.1 <- ggplot(my.data,aes(x = Country, y = Weight_change, color = Diet )) +  geom_point();
gg.base.1 +  stat_summary(aes(group = Diet, color = Diet), geom = "line", fun.y = mean, size = 1.5);
```




```
# The plots suggest that the effect of diet
# is not consistent across all three countries.
# While Diet C showed the greatest mean weight gain
# for USA and UK, for NZ it has a lower mean than Diet A.



mod1 <- lm(Weight_change ~ Country + Diet, data = my.data)
summary(mod1)

##
## Call:
## lm(formula = Weight_change ~ Country + Diet, data = my.data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.025375 -0.010729 -0.001458  0.011250  0.021250
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.126458   0.006185  20.445 2.99e-16 ***
```

```
## CountryUK   -0.070667   0.006797 -10.397 3.65e-10 ***
## CountryNZ   -0.057708   0.006636  -8.697 9.94e-09 ***
## DietB       -0.025000   0.006797  -3.678  0.00125 **
## DietC        0.006625   0.006636   0.998  0.32848
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01442 on 23 degrees of freedom
## Multiple R-squared:  0.8644,Adjusted R-squared:  0.8408
## F-statistic: 36.66 on 4 and 23 DF,  p-value: 1.146e-09
```

```
mod2 <- lm(Weight_change ~ Country * Diet, data = my.data)
summary(mod2)
```

```
##
## Call:
## lm(formula = Weight_change ~ Country * Diet, data = my.data)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0083333 -0.0055000  0.0008333  0.0046667  0.0090000
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.119000   0.003664  32.476  < 2e-16 ***
## CountryUK        -0.068667   0.005182 -13.251 4.77e-11 ***
## CountryNZ        -0.037333   0.005182  -7.204 7.66e-07 ***
## DietB            -0.024000   0.005182  -4.631 0.000182 ***
## DietC             0.028000   0.005182   5.403 3.26e-05 ***
## CountryUK:DietB -0.002000   0.007329  -0.273 0.787870
## CountryNZ:DietB -0.001000   0.007329  -0.136 0.892899
## CountryUK:DietC -0.004000   0.007329  -0.546 0.591547
## CountryNZ:DietC -0.053667   0.007096  -7.563 3.82e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.006347 on 19 degrees of freedom
## Multiple R-squared:  0.9783,Adjusted R-squared:  0.9692
## F-statistic: 107.1 on 8 and 19 DF,  p-value: 3.725e-14
```

```
anova(mod1,mod2)
```

```
## Analysis of Variance Table
##
## Model 1: Weight_change ~ Country + Diet
## Model 2: Weight_change ~ Country * Diet
##   Res.Df       RSS Df Sum of Sq      F    Pr(>F)
## 1     23 0.0047815
## 2     19 0.0007653  4 0.0040162 24.926 2.477e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
library(emmeans)
```

```r
em0.1 <- emmeans(mod1, pairwise ~ Country)
em0.1$contrasts

##  contrast estimate      SE df t.ratio p.value
##  USA - UK   0.0707 0.00680 23  10.397  <.0001
##  USA - NZ   0.0577 0.00664 23   8.697  <.0001
##  UK - NZ   -0.0130 0.00664 23  -1.953  0.1468
##
## Results are averaged over the levels of: Diet
## P value adjustment: tukey method for comparing a family of 3 estimates

em0.2 <- emmeans(mod2, pairwise ~ Country)

## NOTE: Results may be misleading due to involvement in interactions

em0.2$contrasts

##  contrast estimate      SE df t.ratio p.value
##  USA - UK   0.0707 0.00299 19  23.620  <.0001
##  USA - NZ   0.0556 0.00293 19  18.968  <.0001
##  UK - NZ   -0.0151 0.00293 19  -5.159  0.0002
##
## Results are averaged over the levels of: Diet
## P value adjustment: tukey method for comparing a family of 3 estimates

em1 <- emmeans(mod2, pairwise ~ Country*Diet);
em1$contrasts;

##  contrast       estimate      SE df t.ratio p.value
##  USA A - UK A   0.068667 0.00518 19  13.251  <.0001
##  USA A - NZ A   0.037333 0.00518 19   7.204  <.0001
##  USA A - USA B  0.024000 0.00518 19   4.631  0.0045
##  USA A - UK B   0.094667 0.00518 19  18.268  <.0001
##  USA A - NZ B   0.062333 0.00518 19  12.029  <.0001
##  USA A - USA C -0.028000 0.00518 19  -5.403  0.0009
##  USA A - UK C   0.044667 0.00518 19   8.619  <.0001
##  USA A - NZ C   0.063000 0.00485 19  12.997  <.0001
##  UK A - NZ A   -0.031333 0.00518 19  -6.046  0.0002
##  UK A - USA B  -0.044667 0.00518 19  -8.619  <.0001
##  UK A - UK B    0.026000 0.00518 19   5.017  0.0020
##  UK A - NZ B   -0.006333 0.00518 19  -1.222  0.9414
##  UK A - USA C  -0.096667 0.00518 19 -18.654  <.0001
##  UK A - UK C   -0.024000 0.00518 19  -4.631  0.0045
##  UK A - NZ C   -0.005667 0.00485 19  -1.169  0.9539
##  NZ A - USA B  -0.013333 0.00518 19  -2.573  0.2600
##  NZ A - UK B    0.057333 0.00518 19  11.064  <.0001
##  NZ A - NZ B    0.025000 0.00518 19   4.824  0.0030
##  NZ A - USA C  -0.065333 0.00518 19 -12.608  <.0001
##  NZ A - UK C    0.007333 0.00518 19   1.415  0.8786
##  NZ A - NZ C    0.025667 0.00485 19   5.295  0.0011
##  USA B - UK B   0.070667 0.00518 19  13.637  <.0001
##  USA B - NZ B   0.038333 0.00518 19   7.397  <.0001
##  USA B - USA C -0.052000 0.00518 19 -10.035  <.0001
```

```
##  USA B - UK C   0.020667 0.00518 19   3.988  0.0177
##  USA B - NZ C   0.039000 0.00485 19   8.046  <.0001
##  UK B - NZ B   -0.032333 0.00518 19  -6.239  0.0002
##  UK B - USA C  -0.122667 0.00518 19 -23.671  <.0001
##  UK B - UK C   -0.050000 0.00518 19  -9.649  <.0001
##  UK B - NZ C   -0.031667 0.00485 19  -6.533  0.0001
##  NZ B - USA C  -0.090333 0.00518 19 -17.432  <.0001
##  NZ B - UK C   -0.017667 0.00518 19  -3.409  0.0577
##  NZ B - NZ C    0.000667 0.00485 19   0.138  1.0000
##  USA C - UK C   0.072667 0.00518 19  14.023  <.0001
##  USA C - NZ C   0.091000 0.00485 19  18.773  <.0001
##  UK C - NZ C    0.018333 0.00485 19   3.782  0.0272
##
## P value adjustment: tukey method for comparing a family of 9 estimates
```

```
em2 <- emmeans(mod2, pairwise ~ Diet|Country)
em2$contrasts
```

```
## Country = USA:
##  contrast  estimate      SE df t.ratio p.value
##  A - B     0.024000 0.00518 19   4.631  0.0005
##  A - C    -0.028000 0.00518 19  -5.403  0.0001
##  B - C    -0.052000 0.00518 19 -10.035  <.0001
##
## Country = UK:
##  contrast  estimate      SE df t.ratio p.value
##  A - B     0.026000 0.00518 19   5.017  0.0002
##  A - C    -0.024000 0.00518 19  -4.631  0.0005
##  B - C    -0.050000 0.00518 19  -9.649  <.0001
##
## Country = NZ:
##  contrast  estimate      SE df t.ratio p.value
##  A - B     0.025000 0.00518 19   4.824  0.0003
##  A - C     0.025667 0.00485 19   5.295  0.0001
##  B - C     0.000667 0.00485 19   0.138  0.9896
##
## P value adjustment: tukey method for comparing a family of 3 estimates
```

```
em3 <- emmeans(mod2, pairwise ~ Country|Diet)
em3$contrasts
```

```
## Diet = A:
##  contrast estimate      SE df t.ratio p.value
##  USA - UK   0.0687 0.00518 19  13.251  <.0001
##  USA - NZ   0.0373 0.00518 19   7.204  <.0001
##  UK - NZ   -0.0313 0.00518 19  -6.046  <.0001
##
## Diet = B:
##  contrast estimate      SE df t.ratio p.value
##  USA - UK   0.0707 0.00518 19  13.637  <.0001
##  USA - NZ   0.0383 0.00518 19   7.397  <.0001
##  UK - NZ   -0.0323 0.00518 19  -6.239  <.0001
```

```
##
## Diet = C:
##  contrast estimate      SE df t.ratio p.value
##  USA - UK   0.0727 0.00518 19  14.023  <.0001
##  USA - NZ   0.0910 0.00485 19  18.773  <.0001
##  UK - NZ    0.0183 0.00485 19   3.782  0.0034
##
## P value adjustment: tukey method for comparing a family of 3 estimates
```

3 A study on nutrition education

Q3.1: The example is taken from the website: `https://rcompanion.org/handbook/I_09.html`. The data set consists of measurements obtained from a (mock) study of the effect of nutritional education on calorie consumption. The students enrolled in this experiment were randomly assigned to one of three groups each receiving instruction in nutrition education. The students were then asked to document their daily calorie intake once a month for six months.

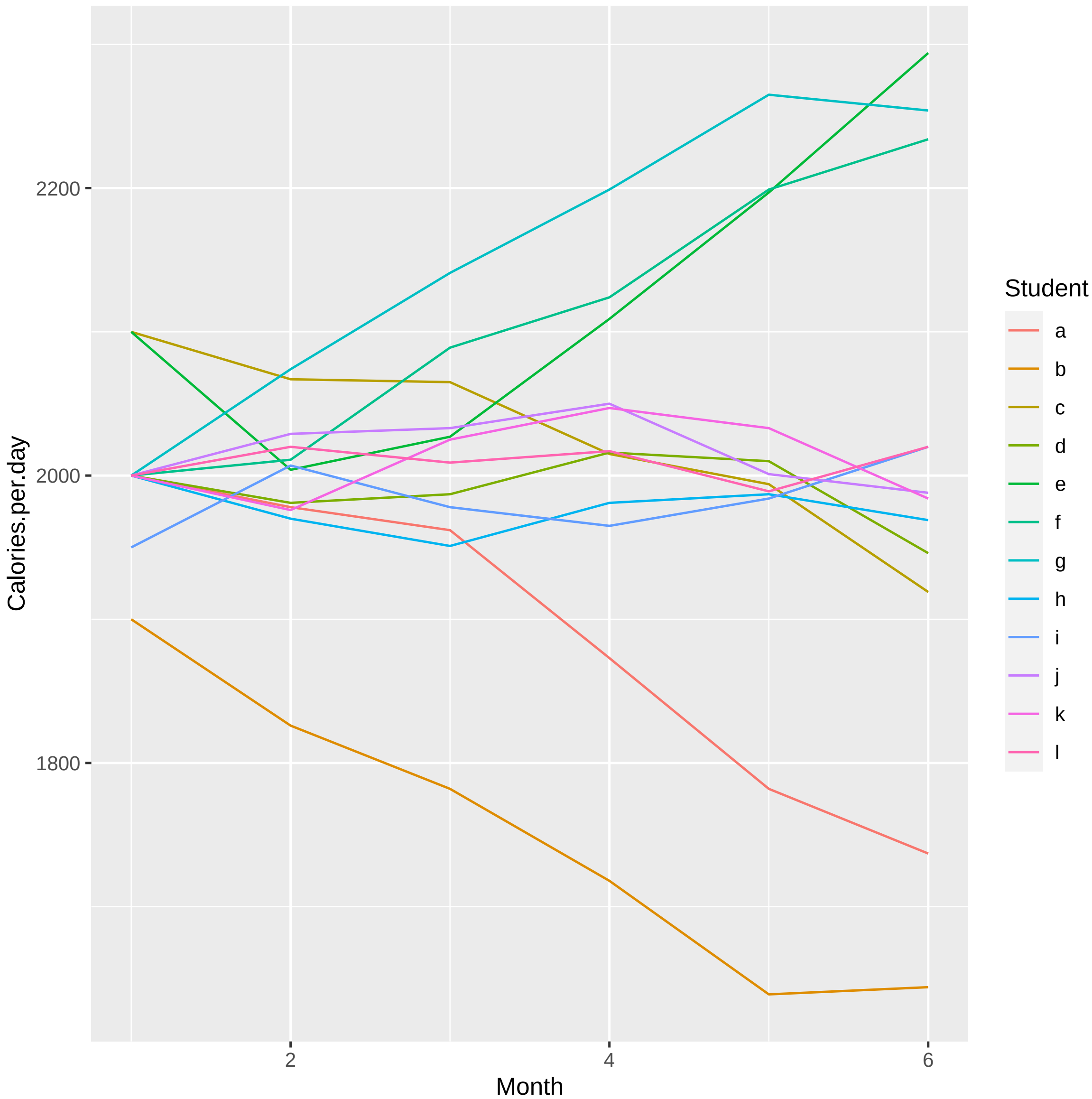
(a) Download the data set in the file `"InstructionCalories.txt"`.

(b) Using an appropriate statistical method, analyse the data to address the research question: which curriculum is better at decreasing calorie intake in students?

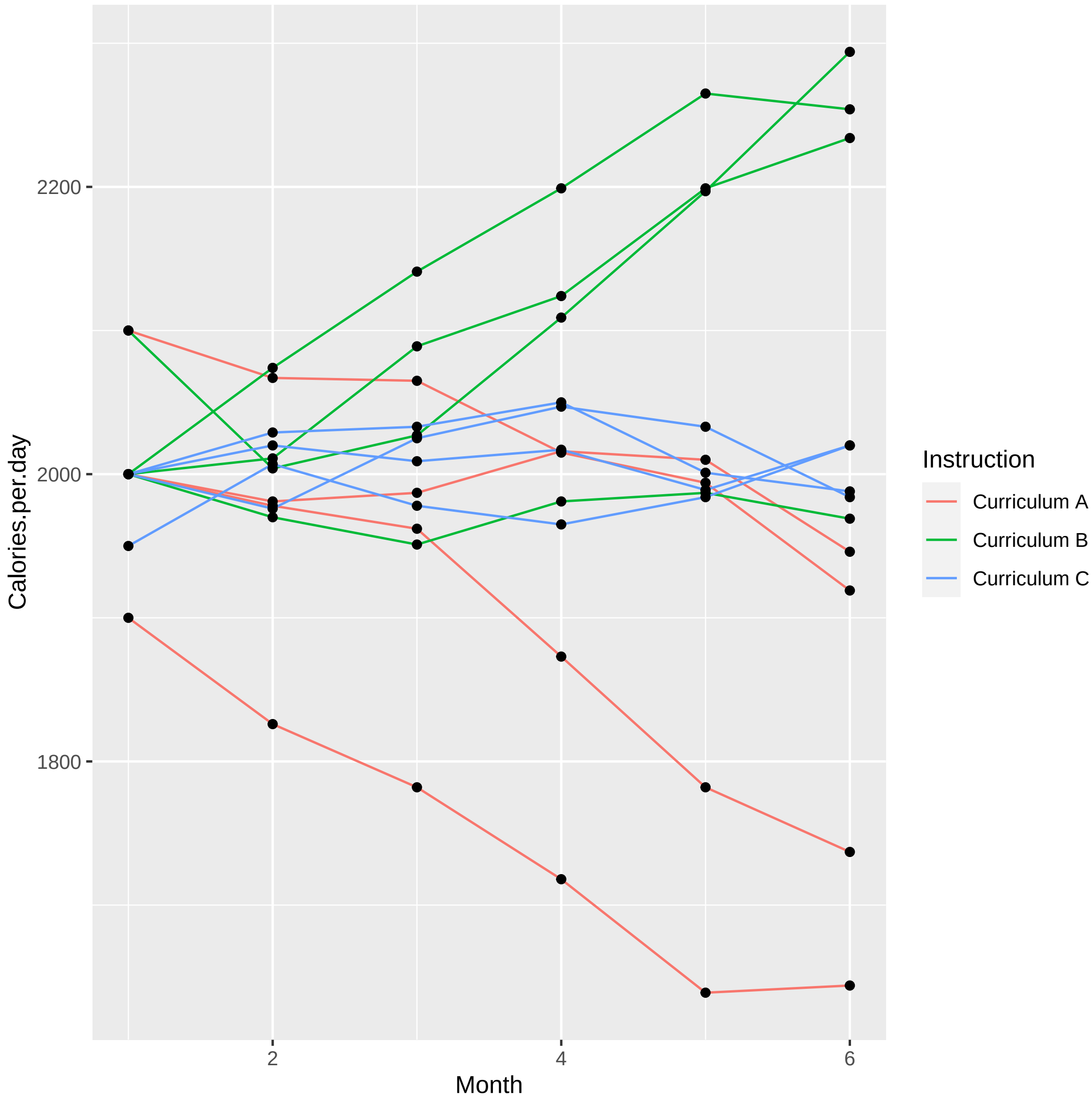
Hint: A hierarchical mixed-effects model with random intercept would be a good tool to analyse these data.

```
##
## How to approach:
## 1. Analyze the difference between each student's final and initial intake. (One-way ANOVA)
## 2. Use all measurements and apply repeated measures ANOVA
## In the second case, we can also specify the  autocorrelation structure.


Input <- ([3006 chars quoted with '"'])

Data <- read.table(textConnection(Input),header=TRUE)

## Or read from a data file:
Data <- read.table("data/InstructionCalories.txt", sep = ",", header = TRUE)


###  Order factors by the order in data frame
###  Otherwise, R will alphabetize them
Data$Instruction <- factor(Data$Instruction,
                         levels=unique(Data$Instruction))


library(ggplot2)

gg.base <- ggplot(Data,aes(x = Month, y = Calories.per.day))
```

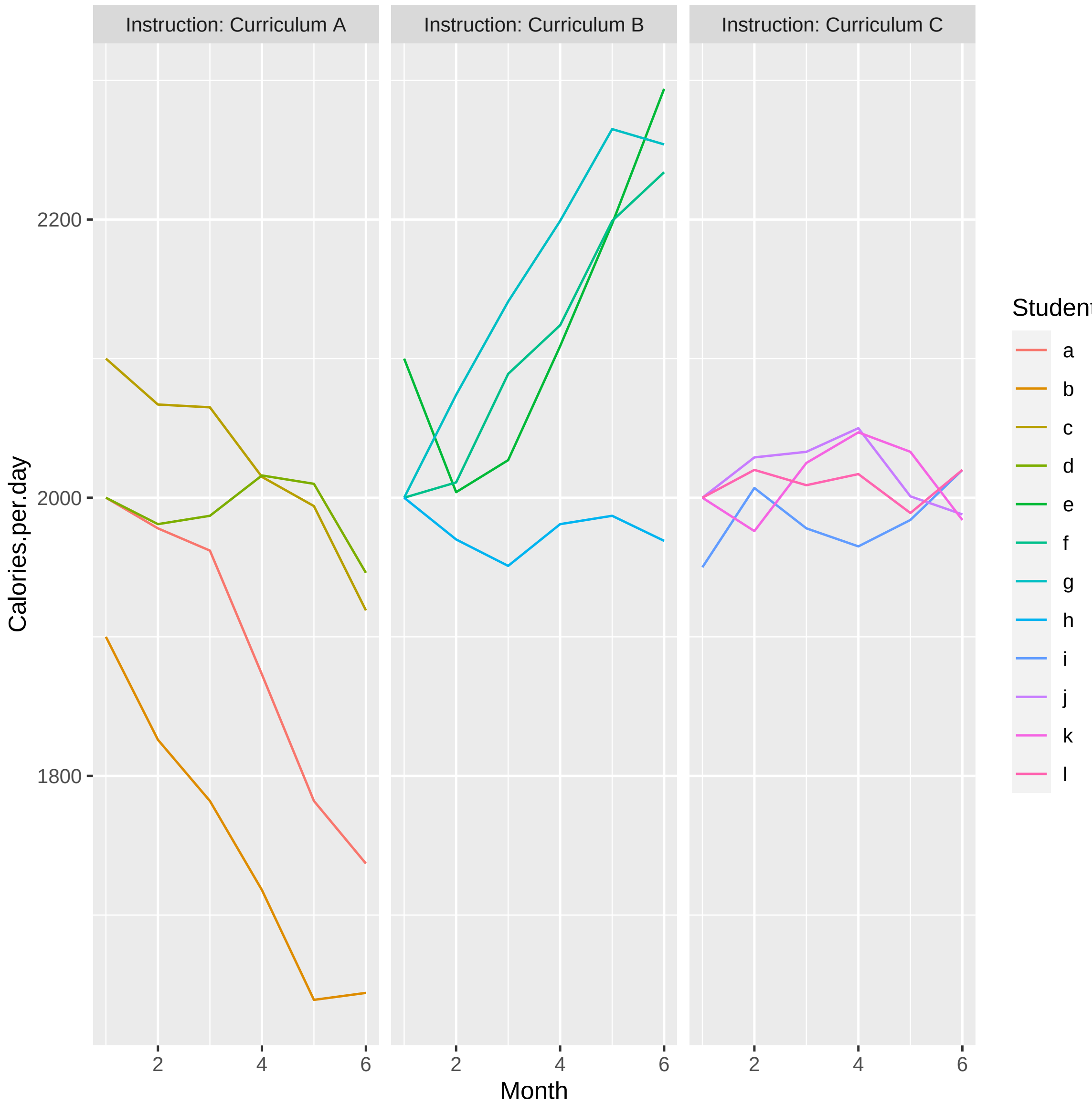
```
gg.idline <- gg.base + geom_line(aes(color = Student, group = Student))
gg.idline
```




```
gg.Gline <- gg.base + geom_line(aes(color = Instruction, group = Student))
gg.Gline +  geom_point()
```




```
gg.idline + facet_wrap( ~ Instruction, labeller = label_both)
```




```
library(nlme)
library(emmeans)

model <- lme(Calories.per.day ~ Instruction + Month + Instruction*Month,
            random = ~1|Student,
            correlation = corAR1(form = ~ Month | Student,
                                 value = 0.4287),
            data=Data,
            method="REML")

summary(model)

## Linear mixed-effects model fit by REML
##   Data: Data
```

```
##        AIC      BIC    logLik
##   716.9693 736.6762 -349.4847
##
## Random effects:
##  Formula: ~1 | Student
##         (Intercept) Residual
## StdDev:  0.02681555 94.74736
##
## Correlation Structure: AR(1)
##  Formula: ~Month | Student
##  Parameter estimate(s):
##       Phi
## 0.9146008
## Fixed effects:  Calories.per.day ~ Instruction + Month + Instruction * Month
##                                    Value Std.Error DF  t-value p-value
## (Intercept)                    2039.5211  51.25269 57 39.79344  0.0000
## InstructionCurriculum B         -50.8577  72.48226  9 -0.70166  0.5006
## InstructionCurriculum C         -52.8492  72.48226  9 -0.72913  0.4845
## Month                           -37.6915   8.03980 57 -4.68811  0.0000
## InstructionCurriculum B:Month    70.3491  11.37000 57  6.18726  0.0000
## InstructionCurriculum C:Month    40.7650  11.37000 57  3.58531  0.0007
##  Correlation:
##                               (Intr) InstCB InstCC Month  InCB:M
## InstructionCurriculum B       -0.707
## InstructionCurriculum C       -0.707  0.500
## Month                         -0.549  0.388  0.388
## InstructionCurriculum B:Month  0.388 -0.549 -0.275 -0.707
## InstructionCurriculum C:Month  0.388 -0.275 -0.549 -0.707  0.500
##
## Standardized Within-Group Residuals:
##        Min         Q1        Med         Q3        Max
## -2.2756199 -0.2583950  0.1082315  0.5105510  1.6774752
##
## Number of Observations: 72
## Number of Groups: 12
```

```
em <- emmeans(model, pairwise ~ Instruction)
```

```
## NOTE: Results may be misleading due to involvement in interactions
```

```
em$contrasts
```

```
##  contrast                    estimate   SE df t.ratio p.value
##  Curriculum A - Curriculum B   -195.4 60.6  9  -3.225  0.0255
##  Curriculum A - Curriculum C    -89.8 60.6  9  -1.483  0.3433
##  Curriculum B - Curriculum C    105.5 60.6  9   1.742  0.2427
##
## Degrees-of-freedom method: containment
## P value adjustment: tukey method for comparing a family of 3 estimates
```

```
# The random effects in the model can be tested by
# comparing this model with random effects to a model
# fitted with just the fixed effects and excluding the random effects.
```

```r
model.fixed <- gls(Calories.per.day ~ Instruction + Month +
                          Instruction*Month,
                  data=Data,
                  method="REML")

anova(model, model.fixed)
```

```
##             Model df      AIC      BIC    logLik   Test L.Ratio p-value
## model           1  9 716.9693 736.6762 -349.4847
## model.fixed     2  7 813.6213 828.9489 -399.8106 1 vs 2 100.652  <.0001
```