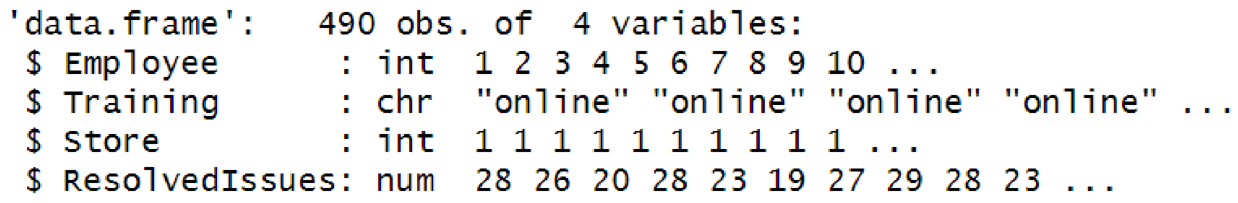
Statistics Assignment 5

- Joshua Boryer 41497475

1a)

Checking the data using the str command in R:



```
'data.frame':	490 obs. of  4 variables:
 $ Employee      : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Training      : chr  "online" "online" "online" "online" ...
 $ Store         : int  1 1 1 1 1 1 1 1 1 1 ...
 $ ResolvedIssues: num  28 26 20 28 23 19 27 29 28 23 ...
```

Everything has loaded correctly.

I observe that

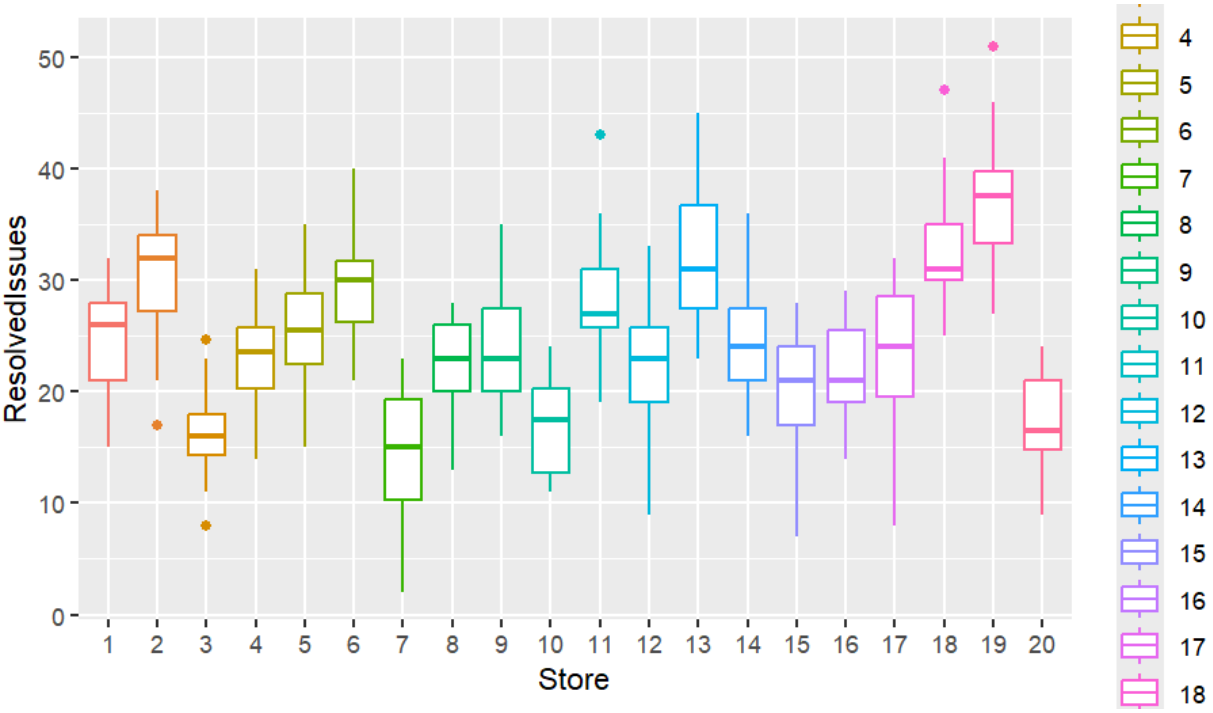
- Employees are labelled from 1 - 490
- Training is a categorical variable of “online” and “in-store”
- Store is a variable holding numbers 1 - 20
- Resolved issues is number containing any integer or decimal number between 1 and 52

1b)

The categorical variables for this dataset are “Training” and “Store” so these should be converted to a factor using the code lines:

```
productivity_2025$Training <- as.factor(productivity_2025$Training)

productivity_2025$Store <- as.factor(productivity_2025$Store)
```

The following ggplot2 graph below shows a boxplot graph comparison of each different store and how many ResolvedIssues they have.



This is a boxplot graph of Store against Resolved Issues showing a relationship between the two variables in whether the store plays a difference to the amount of resolved issues. Currently from this boxplot, the residuals are noted in the {2, 3, 11, 18, 19} for the x axis (Store). In general from this graph, it looks to be that Store 18 has the highest average Resolved Issues but it’s uncertain as to whether that high residual is playing a factor. The lowest using the eye test look to be the Stores {3, 7, 10} with Store 7 showing a large lower quartile range. Overall there’s no consistent pattern here so more in depth data manoeuvres are necessary.

1c)

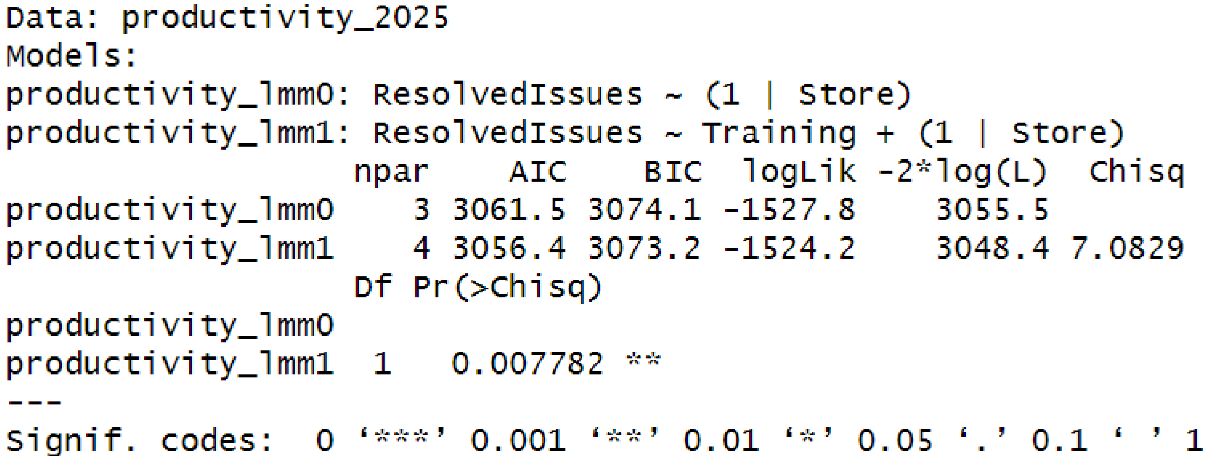
The research question is “Does the mode of training, in-store or online significantly affect the number of customer issues resolved by employees in a retail chain?”

1d)

Response variable: ResolvedIssues
Explanatory variable: Training
Structural component: Store

1e)

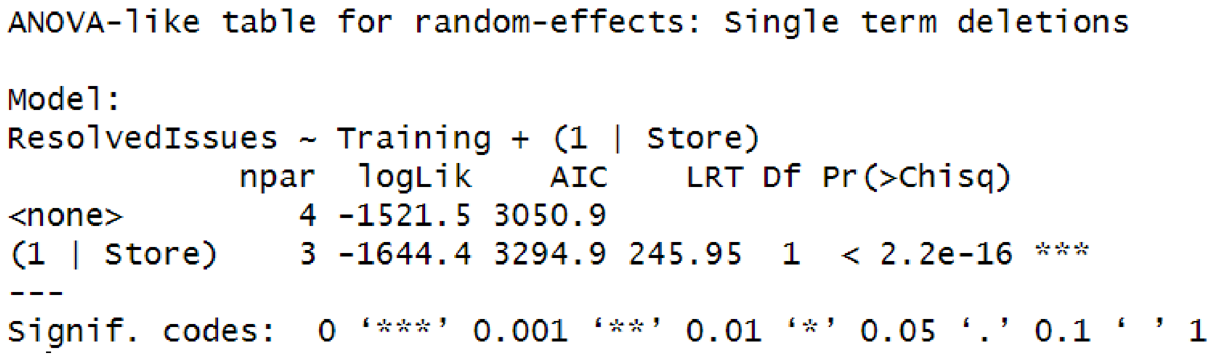
Two linear models lmm0 (Random intercepts model using no slope for training) and lmm1 (Random intercepts model using fixed slope for Training) have been created.
The ANOVA output for the models shows the output:



```
Data: productivity_2025
Models:
productivity_lmm0: ResolvedIssues ~ (1 | Store)
productivity_lmm1: ResolvedIssues ~ Training + (1 | Store)
                  npar    AIC    BIC  logLik -2*log(L)  Chisq
productivity_lmm0    3 3061.5 3074.1 -1527.8    3055.5
productivity_lmm1    4 3056.4 3073.2 -1524.2    3048.4 7.0829
                  Df Pr(>Chisq)
productivity_lmm0
productivity_lmm1  1   0.007782 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

This ANOVA likelihood ratio test output shows that the P-value is 0.007782. P-value < 0.05 therefore the model is significant therefore meaning that adding Training as a fixed effect improves the model, this is also shown by the lower AIC value. Training method has a statistically significant effect on the number of issues resolved.

1f)



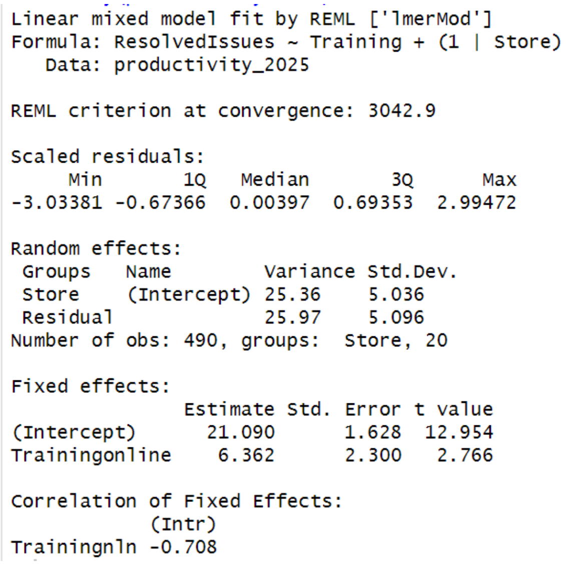
```
ANOVA-like table for random-effects: Single term deletions

Model:
ResolvedIssues ~ Training + (1 | Store)
            npar  logLik    AIC    LRT Df Pr(>Chisq)
<none>         4 -1521.5 3050.9
(1 | Store)    3 -1644.4 3294.9 245.95  1  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

This RANOVA test output provides a P-value of 2.23-16 which shows evidence that the variance in ResolvedIssues is highly significant and that there is a substantial amount of variation in how different stores are resolving issues. This variation is not due to random chance.

There’s significant variability between stores in how they resolve issues.

1g)

Because the better model uses fixed effects of Training the summary of this will be able to answer the research question.



```
Linear mixed model fit by REML ['lmerMod']
Formula: ResolvedIssues ~ Training + (1 | Store)
   Data: productivity_2025

REML criterion at convergence: 3042.9

Scaled residuals:
     Min       1Q   Median       3Q      Max 
-3.03381 -0.67366  0.00397  0.69353  2.99472 

Random effects:
 Groups   Name        Variance Std.Dev.
 Store    (Intercept) 25.36    5.036   
 Residual             25.97    5.096   
Number of obs: 490, groups:  Store, 20

Fixed effects:
               Estimate Std. Error t value
(Intercept)      21.090      1.628  12.954
Trainingonline    6.362      2.300   2.766

Correlation of Fixed Effects:
            (Intr)
Trainingnln -0.708
```

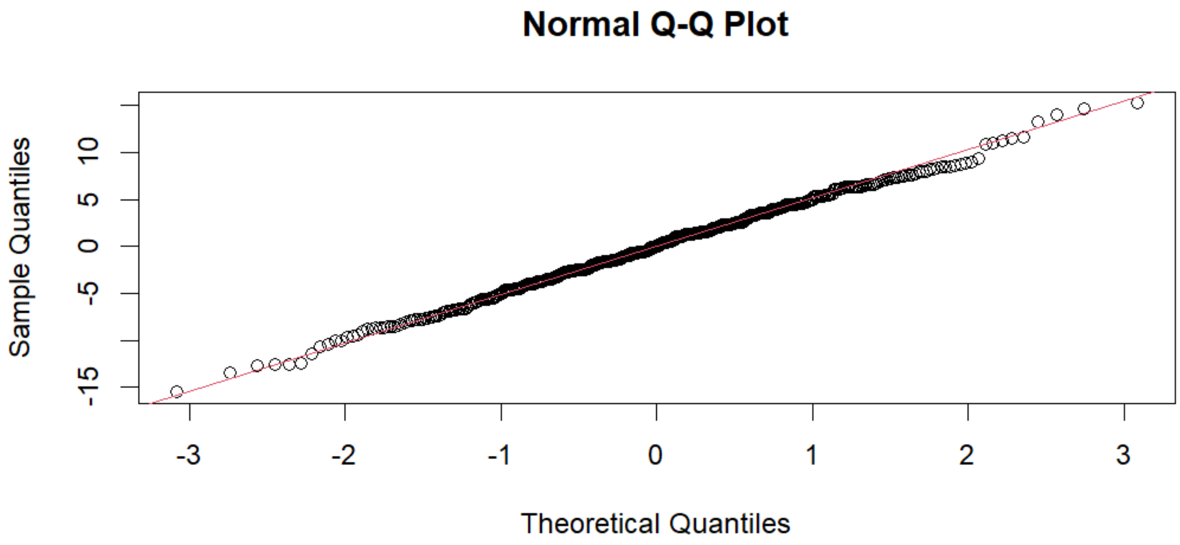
The coefficient for training online is 6.362 (positive).

Question:

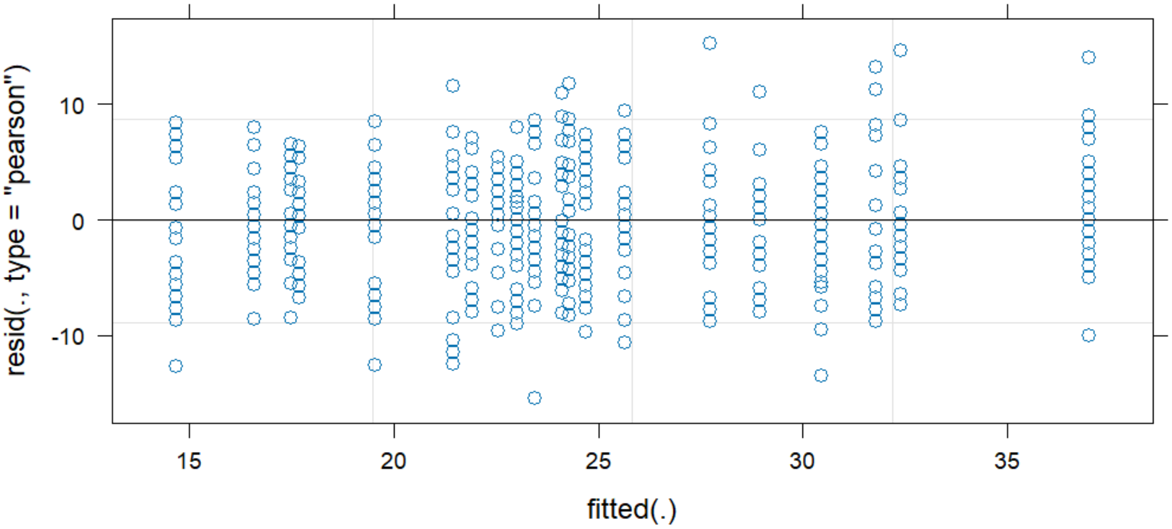
“Does the mode of training, in-store or online significantly affect the number of customer issues resolved by employees in a retail chain?”

Answer: Yes the mode of training, in-store or online does significantly affect the number of customer issues resolved by employees in a retail chain. The evidence shows that the online training mode increases the amount of ResolvedIssues by store on average.

1h)



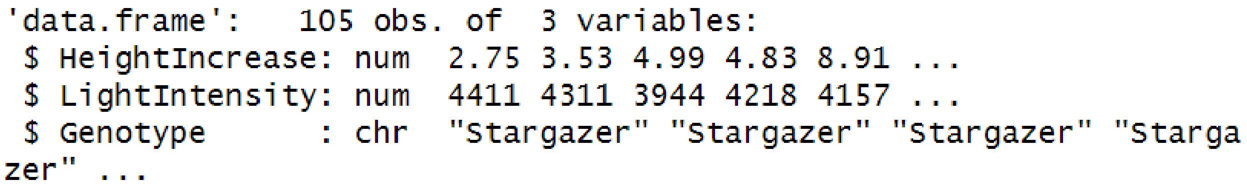
Residuals follow a normal distribution this assumption is met.



Model shows constant variance centered around 0 meeting the homogeneity assumption.

2a)

The str command in R gives these results



```
'data.frame':	105 obs. of  3 variables:
 $ HeightIncrease: num  2.75 3.53 4.99 4.83 8.91 ...
 $ LightIntensity: num  4411 4311 3944 4218 4157 ...
 $ Genotype      : chr  "Stargazer" "Stargazer" "Stargazer" "Starga
zer" ...
```

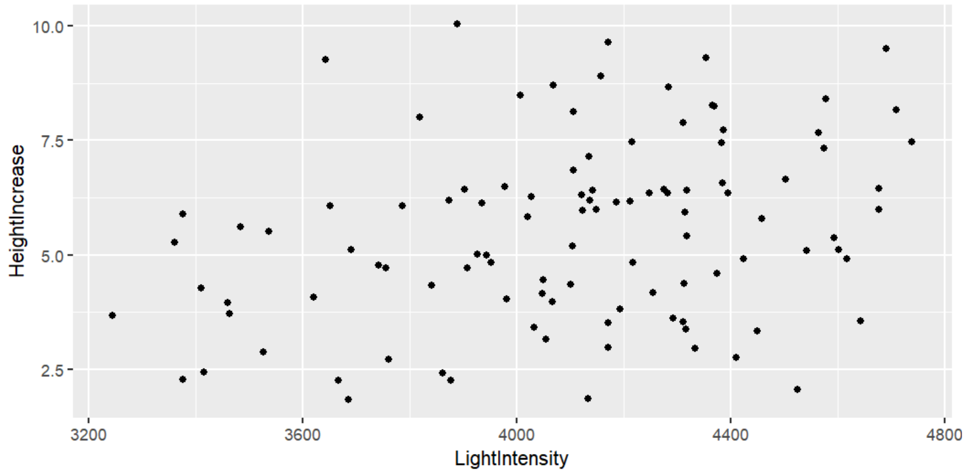
Variables:

- HeightIncrease: Quantitative
- LightIntesity: Quantitative
- Genotype: Categorical

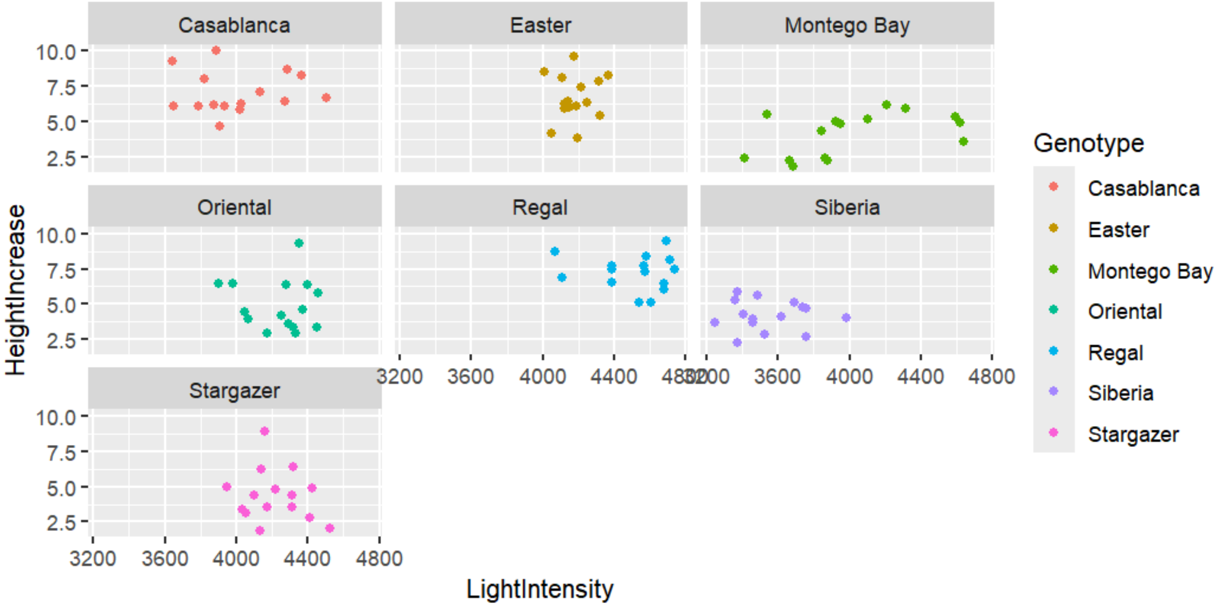
2b)

Research question: “Does LightIntensity have a relationship with HeightIncrease? Do the lilies stay at the same rate height, reduce height, or increase in height when the light intensity changes?

2c)



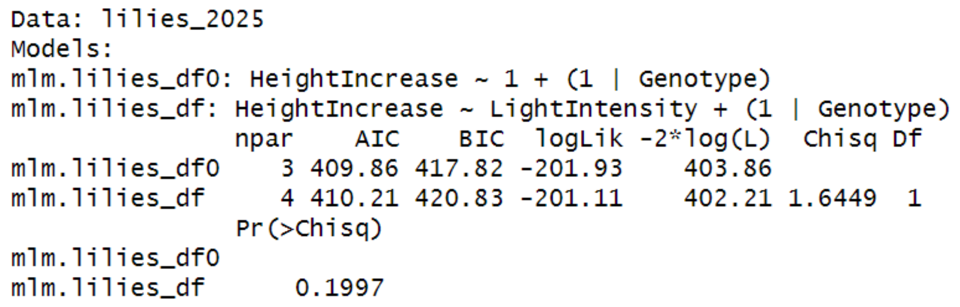
In this ggplot scatter, I observe that there’s no seeable correlation between light intensity and height increase and that there’s no obvious trends or patterns.



This matrix of graphs seperates the variables which makes it easier to see what each genotype is doing. With all of the genotypes noted, there’s still no obvious trends or patterns observed.

2d)

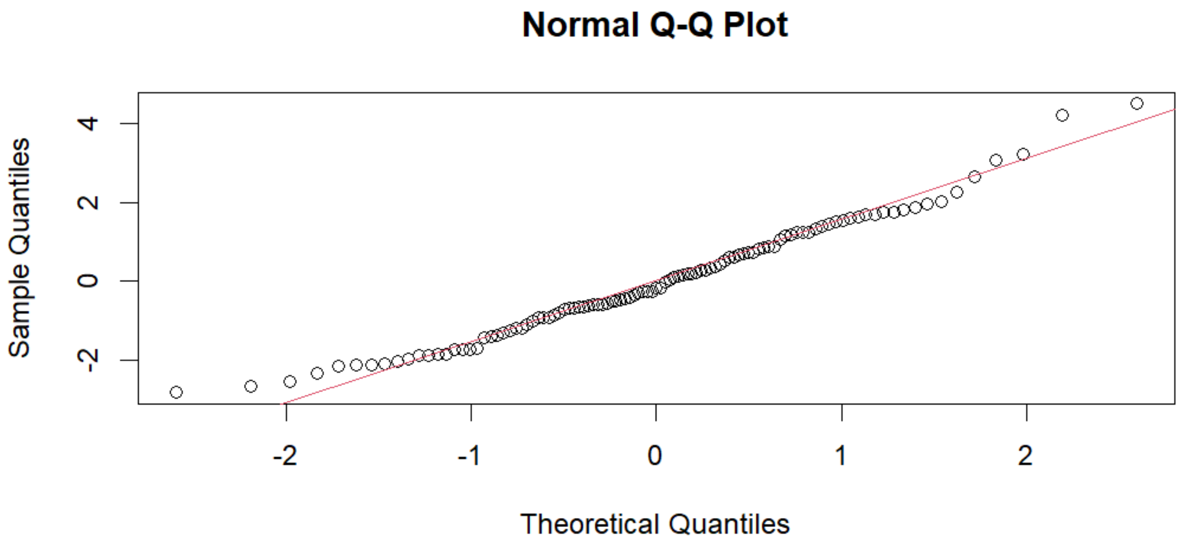
Linear Mixed Models are created and the likelihood test results are below.



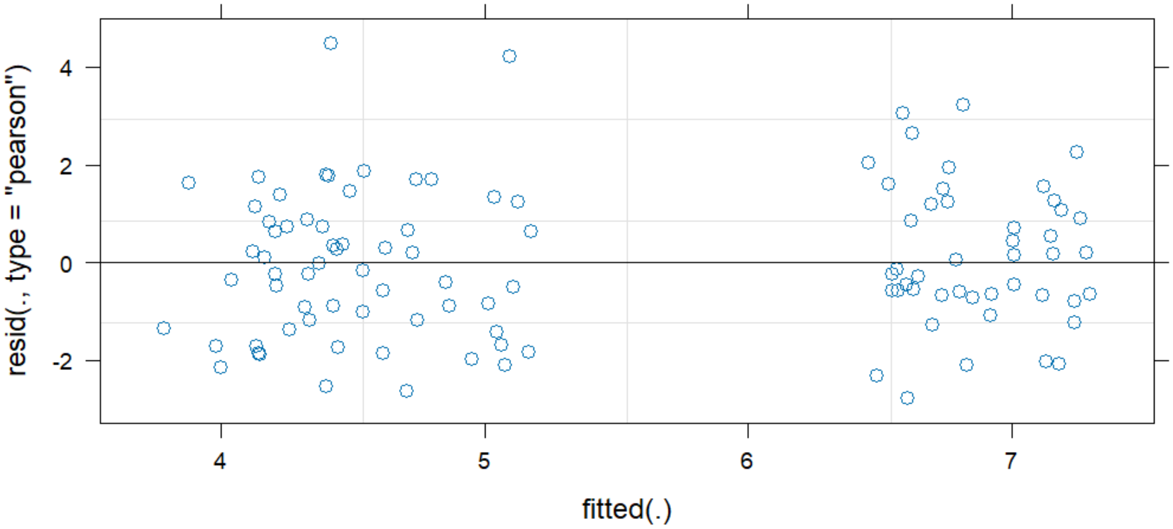
```
Data: lilies_2025
Models:
mlm.lilies_df0: HeightIncrease ~ 1 + (1 | Genotype)
mlm.lilies_df: HeightIncrease ~ LightIntensity + (1 | Genotype)
               npar    AIC    BIC  logLik -2*log(L)  Chisq Df
mlm.lilies_df0    3 409.86 417.82 -201.93    403.86
mlm.lilies_df     4 410.21 420.83 -201.11    402.21 1.6449  1
               Pr(>Chisq)
mlm.lilies_df0
mlm.lilies_df      0.1997
```

Mlm.lilies_df0 has an AIC value of 409.86 and Mlm.lilies_df has an AIC value of 410.21 Meaning that the model without the fixed effect of light intensity is the better fit.
The p-value for mlm.lilies_df of 0.1997 indicates that the addition of light intensity as a fixed effect does not significantly affect the model. There is no strong evidence that light intensity has a significant effect on height increase when accounting for genotype.

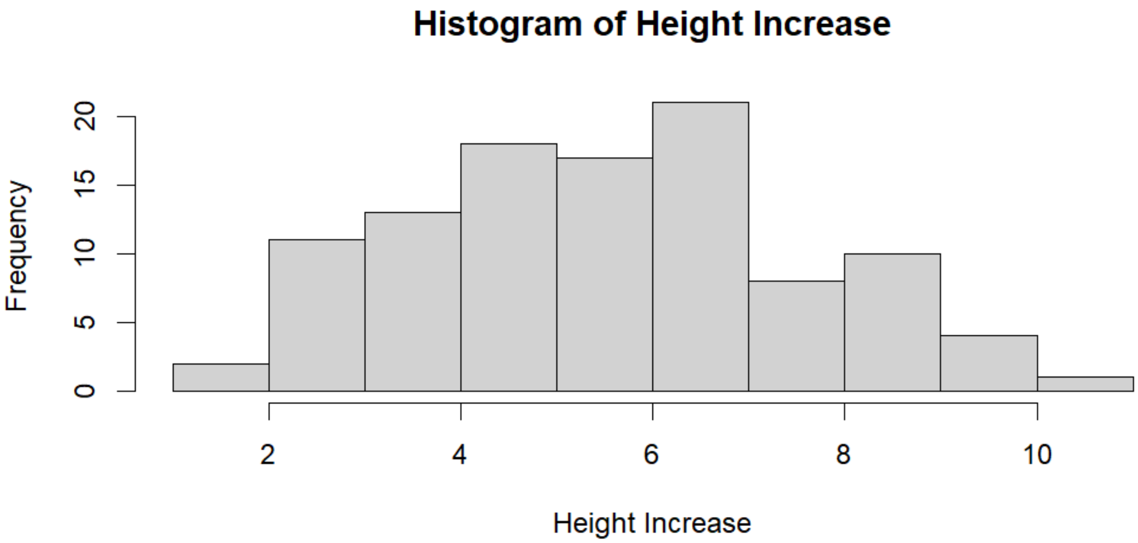
2e)



Residuals follow a normal distribution meeting the assumption of normality.



This graph shows constant variance scattered around 0 meeting the assumption of homogeneity.



Histogram of Height increase is the reponse variable therefore it’s assumed to be a normal distribution which is true in this case.

2f)

In the first question, the fixed effect improved the model seen by the AIC value and P-value in which case we would leave it in the model. For question 2, the fixed effect made no improvement on the model and was not significant meaning there’s no significant relationship between Light Intensity and Height increase for this dataset.