Whitefly Attraction and Disease Dynamics in Cassava Varieties

Koissi Savi

2024-09-26

Background

The experiment aimed to investigate the relationship between cassava variety, whitefly attraction, and Cassava Mosaic Disease (CMD) dynamics. Specifically, it sought to:

1. Determine if distinct Ivorian cassava varieties differentially attract whiteflies.
2. Investigate the correlation between whitefly repellence and CMD resistance.
3. Assess the impact of CMD-resistant or whitefly-repellent cassava varieties on disease dynamics.
4. Examine how virus infection influences whitefly behavior and performance.

Experimental Design

The experiment utilized a Fisher block design with six Ivorian cassava varieties: Yavo (VS), Yacé (VS), Agba blé (S), Bocou 1 (S), Bonoua 34 (R), and TMS30572 (R). Plants were sown 1 meter apart, with 30 plants per plot and 30 m^2 plots. To minimize whitefly movement between plots, a 2-meter distance was maintained between plots.

Data on whitefly presence, abundance, and disease occurrence were collected weekly starting from week 6 until week 62, when plants became less attractive to whiteflies.

Data Analysis

The following sections will detail the data cleaning, exploration, and modeling processes to address the research objectives.

Data Cleaning and Exploration

- **Data Import**: Load the dataset into R and inspect its structure.
- **Data Cleaning**: Identify and address any missing values, outliers, or inconsistencies in the data.

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
##
## Attaching package: 'nlme'
##
```

```
##
## The following object is masked from 'package:dplyr':
##
##     collapse
##
##
## Loading required package: Matrix
##
##
## Attaching package: 'Matrix'
##
##
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
##
##
##
## Attaching package: 'lme4'
##
##
## The following object is masked from 'package:nlme':
##
##     lmList
##
##
## corrplot 0.94 loaded
```

```
## Warning: package 'brms' was built under R version 4.4.1
```

```
## Loading required package: Rcpp
## Loading 'brms' package (version 2.22.0). Useful instructions
## can be found by typing help('brms'). A more detailed introduction
## to the package is available through vignette('brms_overview').
##
## Attaching package: 'brms'
##
## The following object is masked from 'package:glmmTMB':
##
##     lognormal
##
## The following object is masked from 'package:lme4':
##
##     ngrps
##
## The following object is masked from 'package:stats':
##
##     ar
```

```
## # A tibble: 6 x 16
##   block plant  week variety    severity infType nubInfShoots nubOfShoots
##   <dbl> <dbl> <dbl> <chr>         <dbl> <chr>          <dbl>       <dbl>
## 1     1     1     6 Agba blé 3        1 Healthy            0           1
## 2     1     2     6 Agba blé 3        1 Healthy            0           1
## 3     1     3     6 Agba blé 3        1 Healthy            0           1
```

```
## 4     1     4     6 Agba blé 3        1 Healthy            0           1
## 5     1     5     6 Agba blé 3        1 Healthy            0           1
## 6     1     6     6 Agba blé 3        1 Healthy            0           1
## # i 8 more variables: localization <chr>, mosaic <dbl>, leaf_distortion <dbl>,
## #   vein_clearing <lgl>, `chlorotic blotch` <lgl>, filiform <dbl>,
## #   dieback <lgl>, wfCount5leaves <chr>
```

```
##      block          plant           week         variety
##  Min.   : 1.0   Min.   : 1.0   Min.   : 6.00   Length:7560
##  1st Qu.: 5.0   1st Qu.: 8.0   1st Qu.:12.00   Class :character
##  Median : 9.5   Median :15.5   Median :21.00   Mode  :character
##  Mean   : 9.5   Mean   :15.5   Mean   :22.57
##  3rd Qu.:14.0   3rd Qu.:23.0   3rd Qu.:28.00
##  Max.   :18.0   Max.   :30.0   Max.   :62.00
##
##     severity      infType           nubInfShoots     nubOfShoots
##  Min.   :1.00   Length:7560        Min.   : 0.000   Min.   : 1.000
##  1st Qu.:1.00   Class :character   1st Qu.: 0.000   1st Qu.: 1.000
##  Median :1.00   Mode  :character   Median : 0.000   Median : 4.000
##  Mean   :1.63                      Mean   : 1.766   Mean   : 5.953
##  3rd Qu.:3.00                      3rd Qu.: 1.000   3rd Qu.:10.000
##  Max.   :4.00                      Max.   :15.000   Max.   :15.000
##  NA's   :320                       NA's   :914      NA's   :995
##  localization          mosaic     leaf_distortion vein_clearing
##  Length:7560        Min.   :1      Min.   :1       Mode:logical
##  Class :character   1st Qu.:1      1st Qu.:1       NA's:7560
##  Mode  :character   Median :1      Median :1
##                     Mean   :1      Mean   :1
##                     3rd Qu.:1      3rd Qu.:1
##                     Max.   :1      Max.   :1
##                     NA's   :5270   NA's   :5362
##  chlorotic blotch    filiform    dieback        wfCount5leaves
##  Mode:logical     Min.   :1      Mode:logical   Length:7560
##  NA's:7560        1st Qu.:1      NA's:7560      Class :character
##                   Median :1                     Mode  :character
##                   Mean   :1
##                   3rd Qu.:1
##                   Max.   :1
##                   NA's   :6174
```

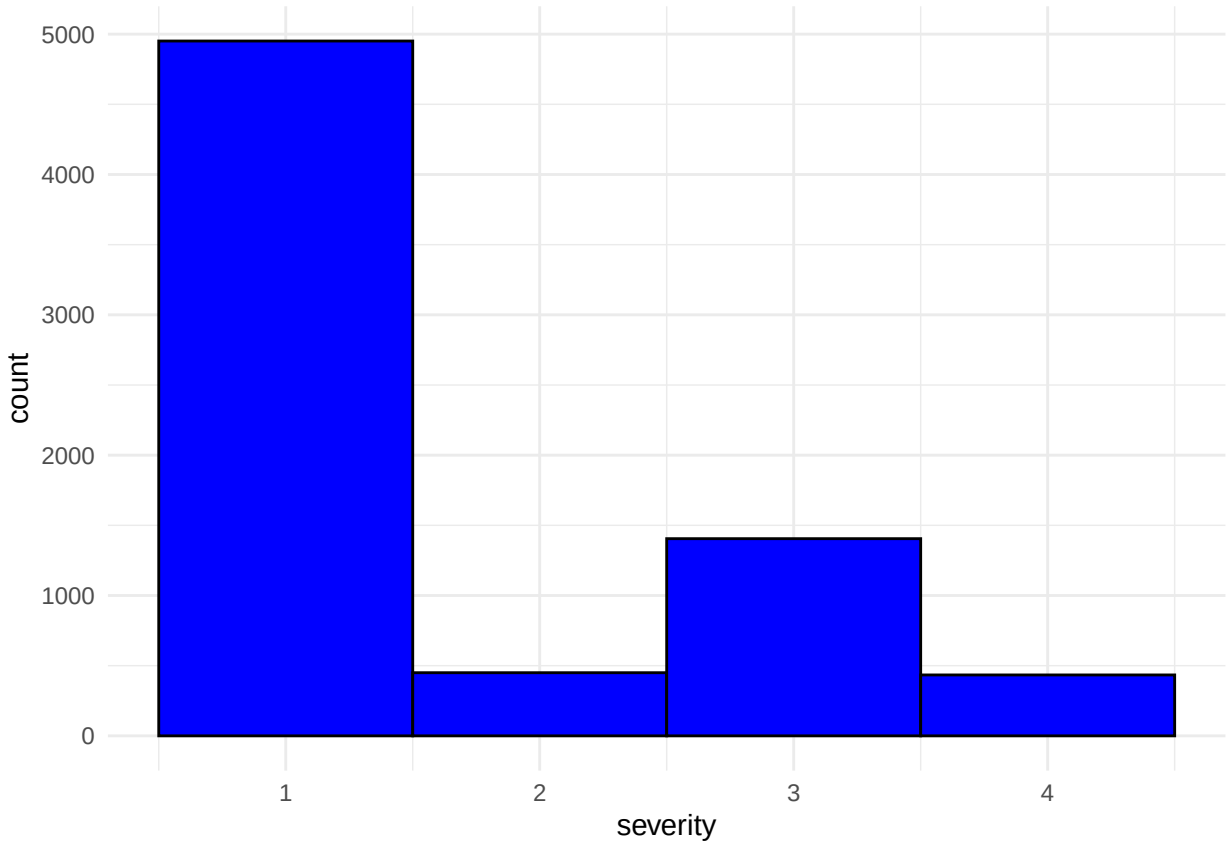
```
## tibble [7,560 x 16] (S3: tbl_df/tbl/data.frame)
##  $ block           : num [1:7560] 1 1 1 1 1 1 1 1 1 1 ...
##  $ plant           : num [1:7560] 1 2 3 4 5 6 7 8 9 10 ...
##  $ week            : num [1:7560] 6 6 6 6 6 6 6 6 6 6 ...
##  $ variety         : chr [1:7560] "Agba blé 3" "Agba blé 3" "Agba blé 3" "Agba blé 3" ...
##  $ severity        : num [1:7560] 1 1 1 1 1 1 1 1 1 1 ...
##  $ infType         : chr [1:7560] "Healthy" "Healthy" "Healthy" "Healthy" ...
##  $ nubInfShoots    : num [1:7560] 0 0 0 0 0 0 0 0 0 0 ...
##  $ nubOfShoots     : num [1:7560] 1 1 1 1 1 1 1 2 1 1 ...
##  $ localization    : chr [1:7560] "No symptoms" "No symptoms" "No symptoms" "No symptoms" ...
##  $ mosaic          : num [1:7560] NA NA NA NA NA NA NA NA NA NA ...
##  $ leaf_distortion : num [1:7560] NA NA NA NA NA NA NA NA NA NA ...
##  $ vein_clearing   : logi [1:7560] NA NA NA NA NA NA ...
##  $ chlorotic blotch: logi [1:7560] NA NA NA NA NA NA ...
##  $ filiform        : num [1:7560] NA NA NA NA NA NA NA NA NA NA ...
```

```
##  $ dieback         : logi [1:7560] NA NA NA NA NA NA ...
##  $ wfCount5leaves  : chr [1:7560] "0" "1" "1" "0" ...
```

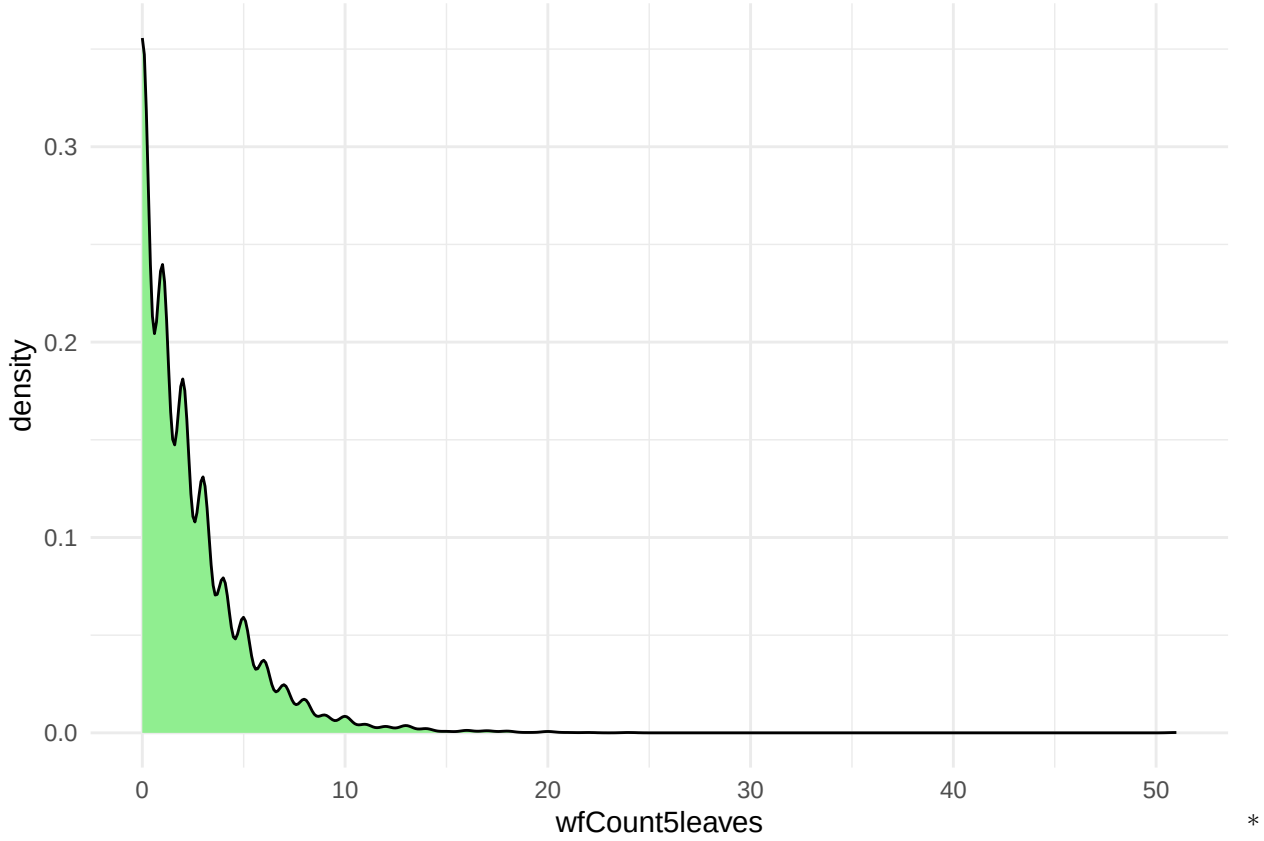
```
##      block            plant            week         variety
##  Min.   : 1.000   Min.   : 1.00   Min.   : 6.00   Length:7242
##  1st Qu.: 5.000   1st Qu.: 8.00   1st Qu.:12.00   Class :character
##  Median : 9.000   Median :16.00   Median :20.00   Mode  :character
##  Mean   : 9.413   Mean   :15.64   Mean   :22.53
##  3rd Qu.:14.000   3rd Qu.:23.00   3rd Qu.:28.00
##  Max.   :18.000   Max.   :30.00   Max.   :62.00
##
##     severity      infType           nubInfShoots     nubOfShoots
##  Min.   :1.00   Length:7242        Min.   : 0.000   Min.   : 1.000
##  1st Qu.:1.00   Class :character   1st Qu.: 0.000   1st Qu.: 1.000
##  Median :1.00   Mode  :character   Median : 0.000   Median : 4.000
##  Mean   :1.63                      Mean   : 1.766   Mean   : 5.953
##  3rd Qu.:3.00                      3rd Qu.: 1.000   3rd Qu.:10.000
##  Max.   :4.00                      Max.   :15.000   Max.   :15.000
##  NA's   :2                         NA's   :596      NA's   :677
##      mosaic       leaf_distortion     filiform      wfCount5leaves
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   : 0.000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.: 0.000
##  Median :0.0000   Median :0.0000   Median :0.0000   Median : 1.000
##  Mean   :0.3162   Mean   :0.3035   Mean   :0.1914   Mean   : 2.202
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.: 3.000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :51.000
##                                                     NA's   :776
##     symptom       sympUpperLeaves  sympLowerLeaves
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :1.0000   Median :0.0000   Median :0.0000
##  Mean   :0.6831   Mean   :0.3085   Mean   :0.2039
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##
```

- **Exploratory Data Analysis (EDA)**: Visualize the distribution of variables, identify patterns, and assess the relationships between variables.
  - Distribution of Numeric Variables

```
## Warning: Removed 2 rows containing non-finite outside the scale range
## (`stat_bin()`).
```

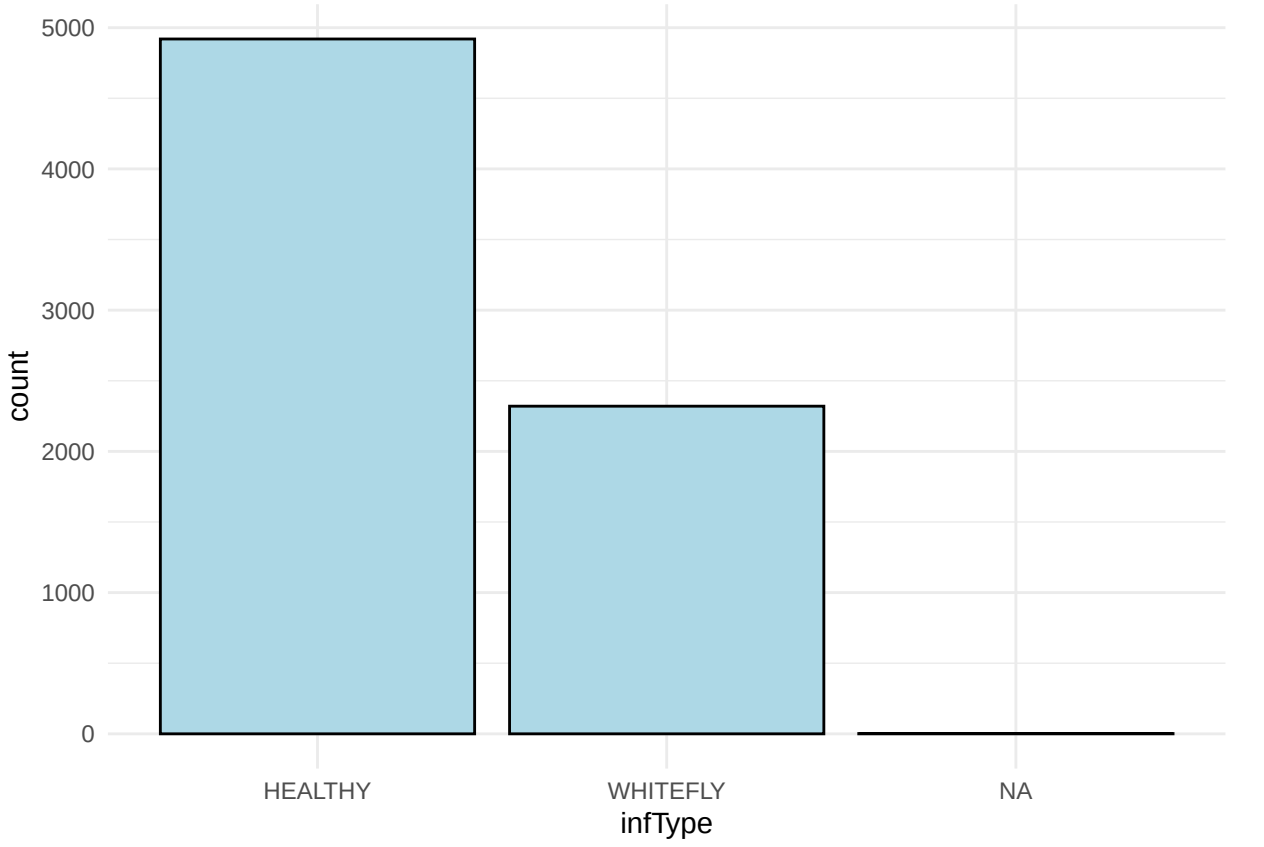


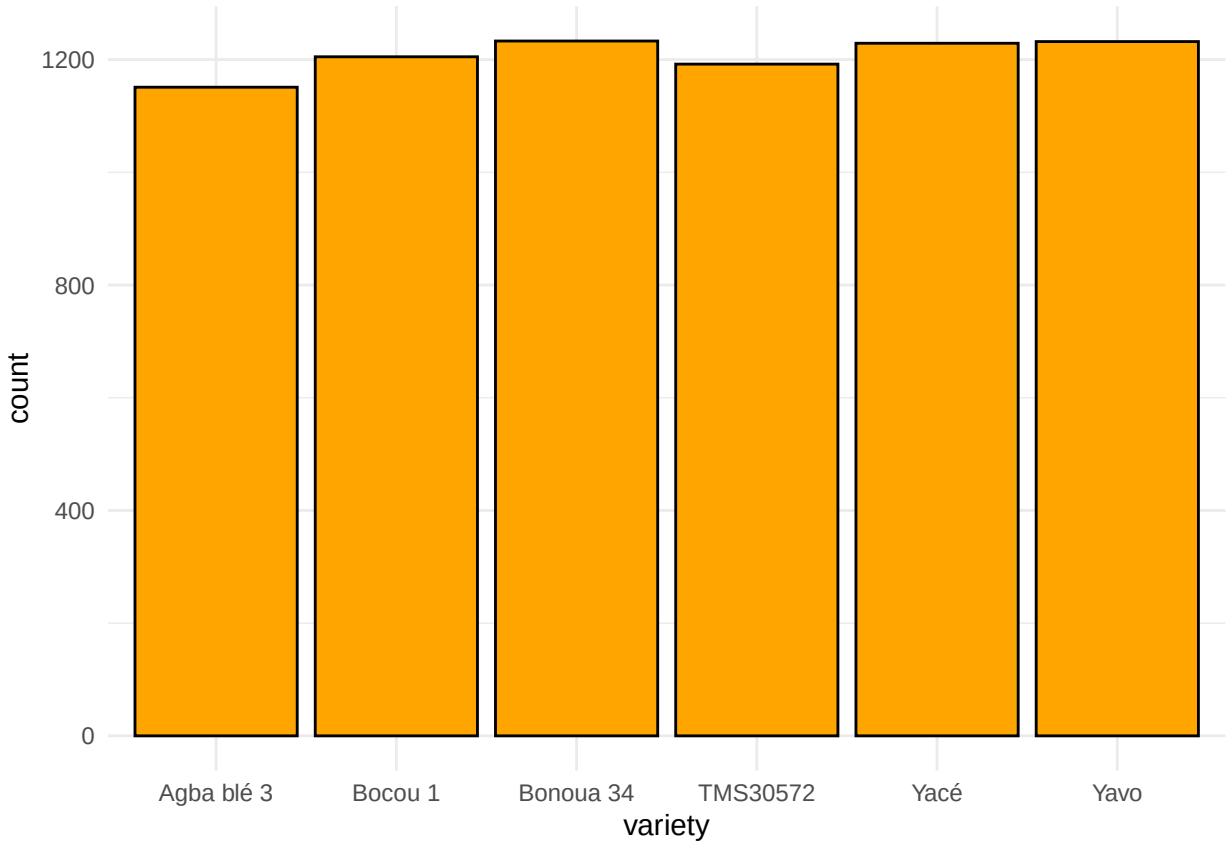

```
## Warning: Removed 776 rows containing non-finite outside the scale range
## (`stat_density()`).
```



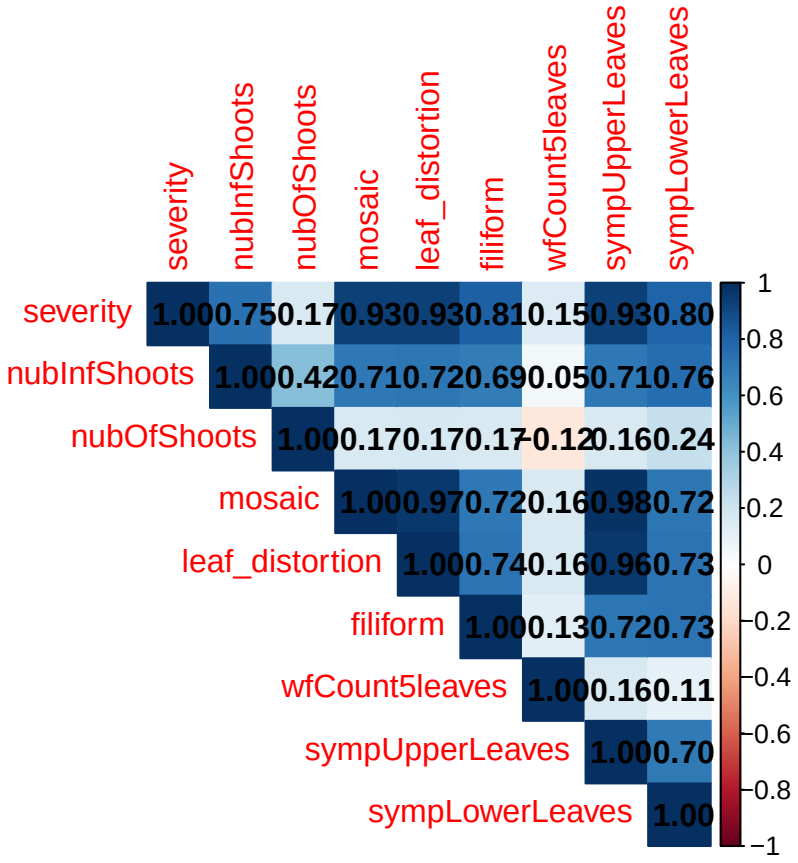
*

Distribution of categorical variables





- Correlation between numeric data

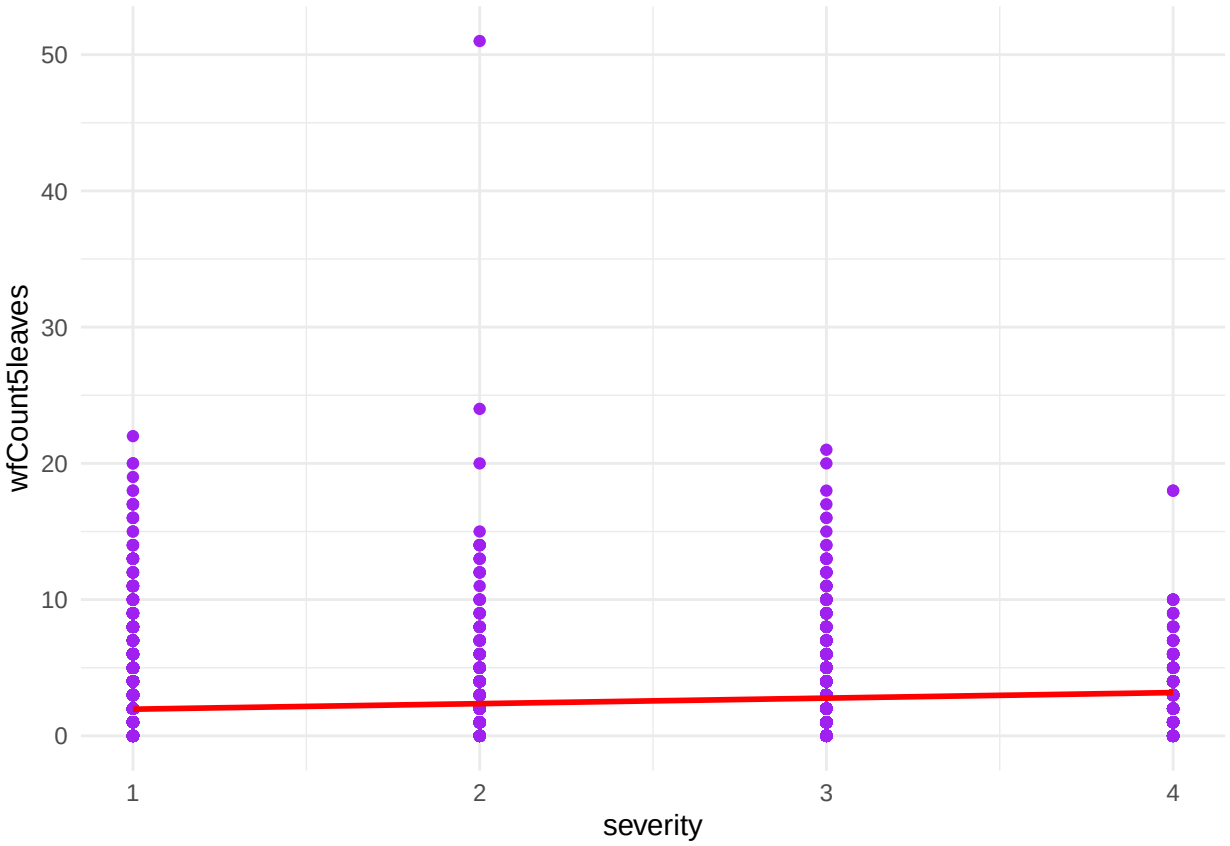


- Relationships Between Variables

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: Removed 776 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

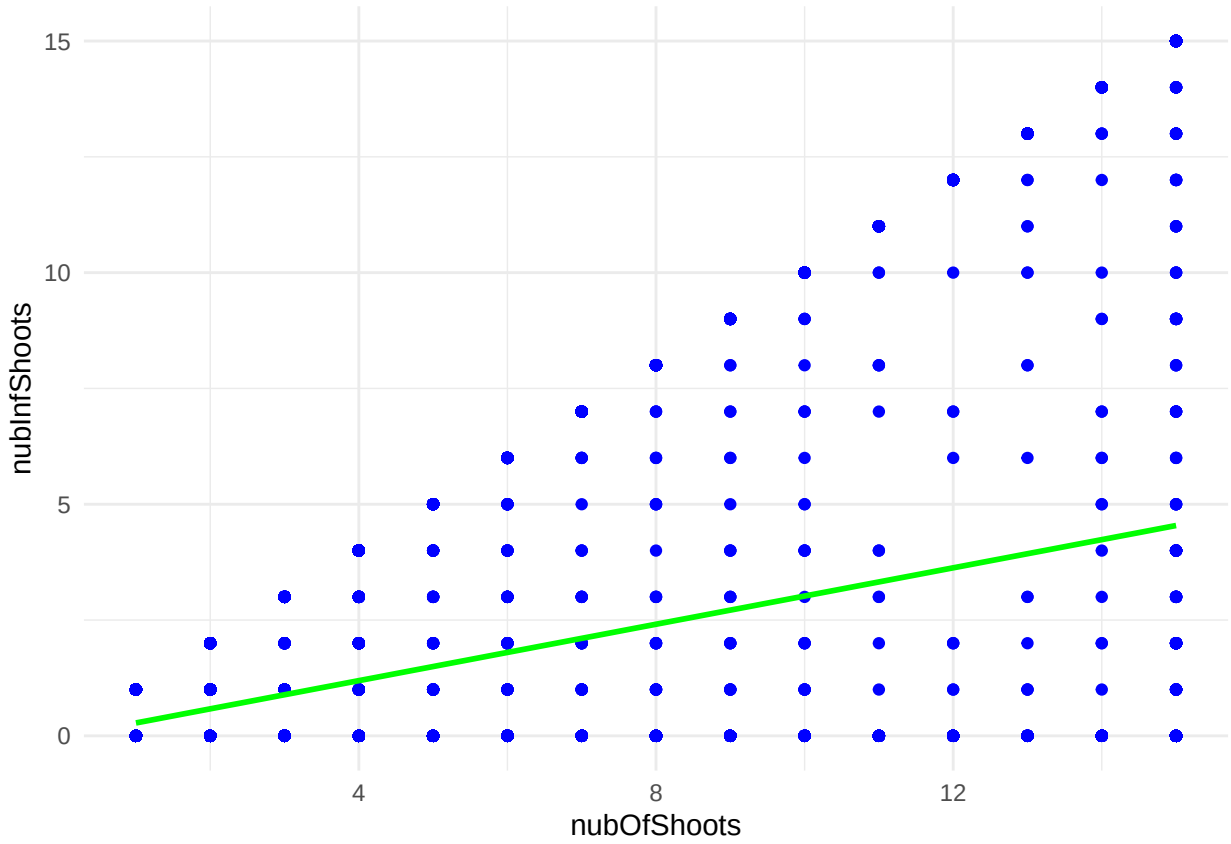
```
## Warning: Removed 776 rows containing missing values or values outside the scale range
## (`geom_point()`).
```




```
## `geom_smooth()` using formula = 'y ~ x'
```

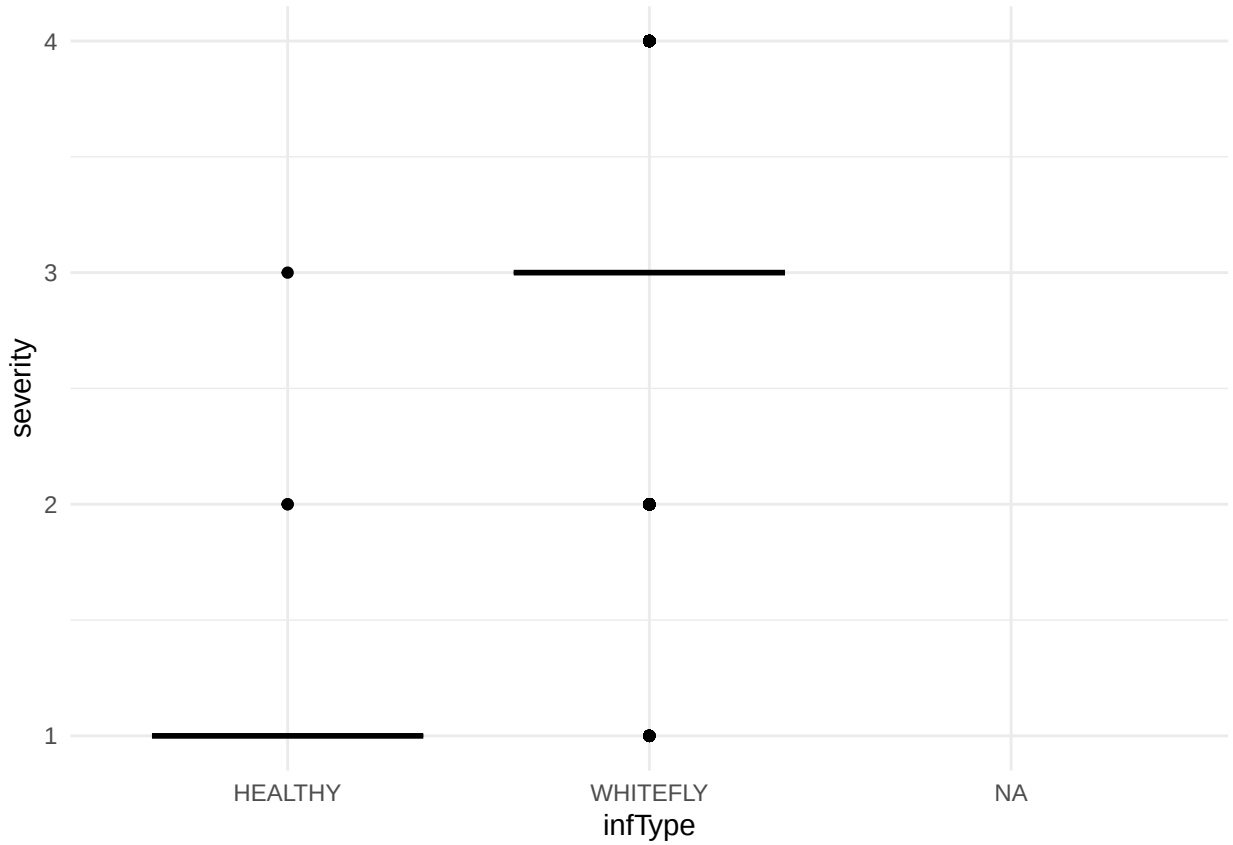
```
## Warning: Removed 678 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

```
## Warning: Removed 678 rows containing missing values or values outside the scale range
## (`geom_point()`).
```

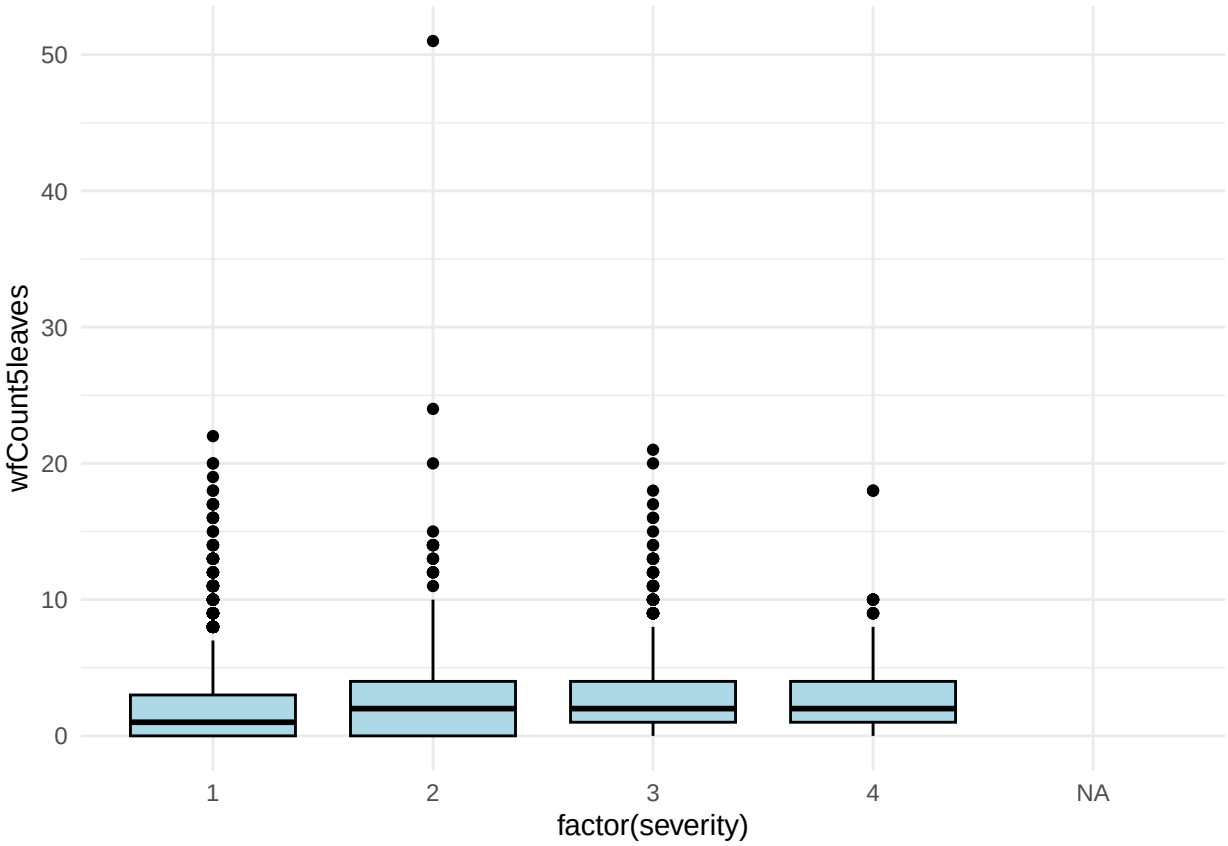


- Boxplots for Numerical Variables

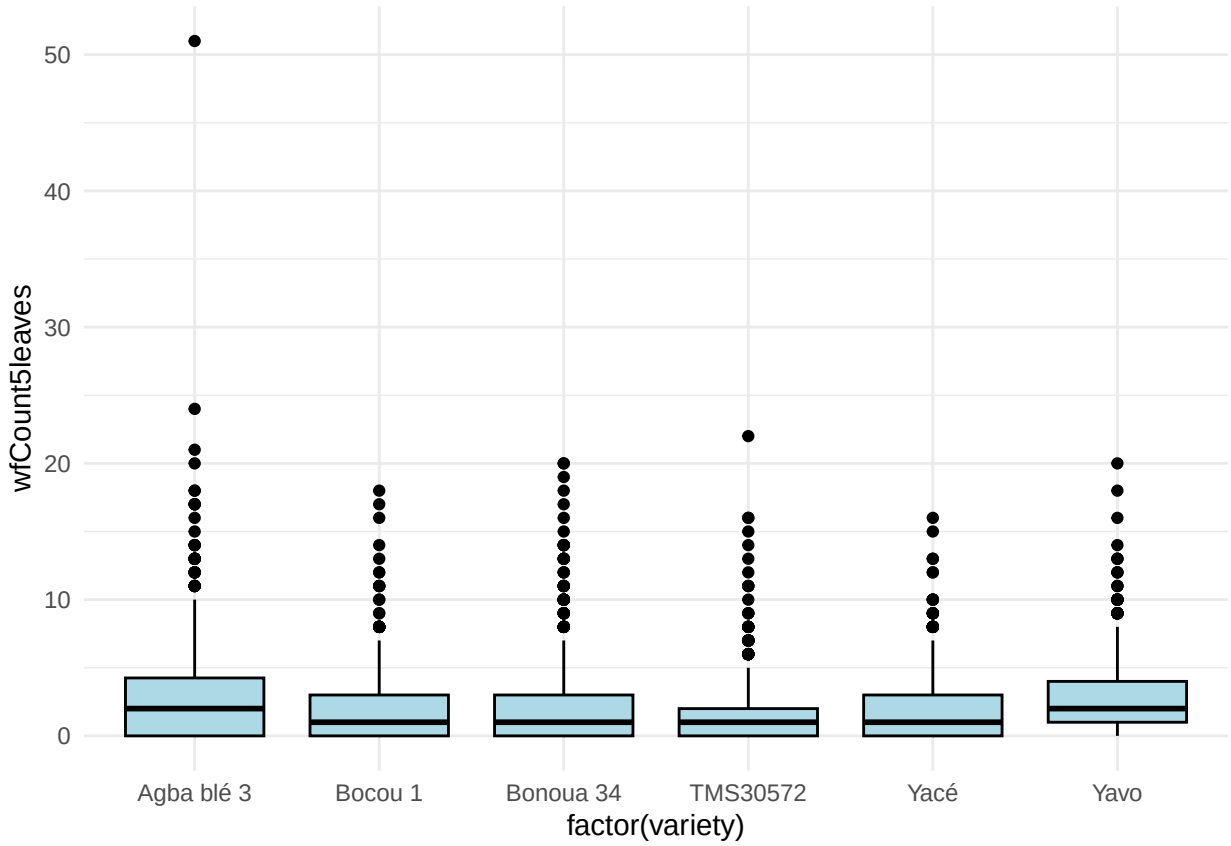
```
## Warning: Removed 2 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```




```
## Warning: Removed 776 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```




```
## Warning: Removed 776 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```



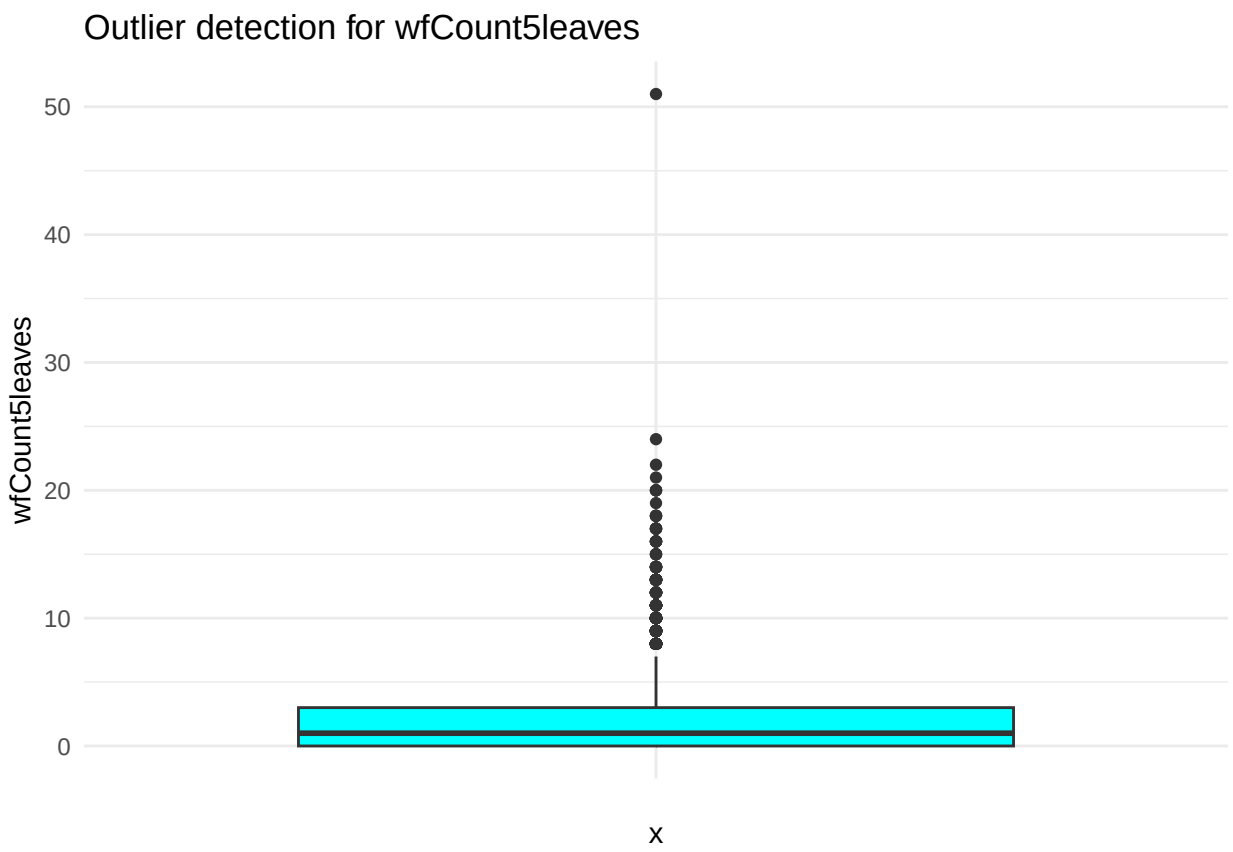
- Summary Table of Categorical Variables

```
## # A tibble: 6 x 2
##   variety    count
##   <chr>      <int>
## 1 Agba blé 3  1151
## 2 Bocou 1     1205
## 3 Bonoua 34   1233
## 4 TMS30572    1192
## 5 Yacé        1229
## 6 Yavo        1232
```

- Identify Outliers

```
# Boxplot to check for outliers in wfCount5leaves
ggplot(dataClean, aes(x = "", y = wfCount5leaves)) +
  geom_boxplot(fill = "cyan") +
  theme_minimal() +
  labs(title = "Outlier detection for wfCount5leaves")
```

```
## Warning: Removed 776 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```



- Remove the Outliers

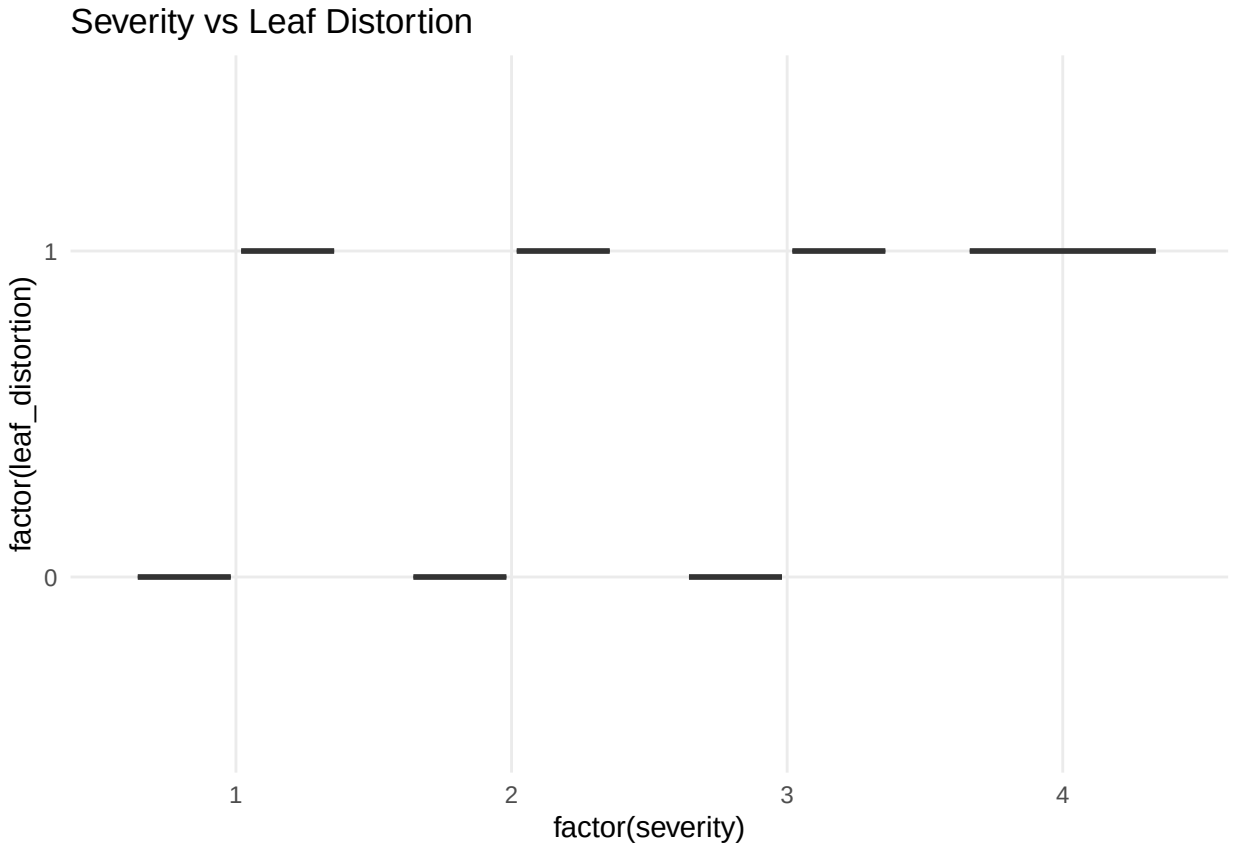
```
dataClean %>% dplyr::select(wfCount5leaves) %>% table()
```

```
## wfCount5leaves
##    0    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15
## 2017 1321 1001  726  435  328  204  136   95   50   47   24   18   21   12    4
##   16   17   18   19   20   21   22   24   51
##    7    6    5    1    4    1    1    1    1
```

```
# Remove rows where wfCount5leaves equals 51
dataClean <- dataClean %>%
  dplyr::filter(wfCount5leaves != 51)
```

- Visualizing Relationships Between Symptoms and Severity

```
# Boxplot to see the relationship between severity and leaf distortion
ggplot(dataClean, aes(x = factor(severity), y = factor(leaf_distortion))) +
  geom_boxplot() +
  theme_minimal() +
  labs(title = "Severity vs Leaf Distortion")
```



Objective 1: Whitefly Attraction

- **Statistical Model**: Use a suitable statistical model generalized linear model to analyze the relationship between cassava variety and whitefly abundance.
- **Hypothesis Testing**: Test the hypothesis that there are significant differences in whitefly attraction among the cassava varieties.

  Model Selection

  - Response Variable: wfCount5leaves (likely count data).
  - Fixed Effect: variety (cassava variety).
  - Random Effects: block, plant, and week.
- **Statistical Model**: Generalized Linear Mixed Model (GLMM)

Since the response variable wfCount5leaves is a count, a Poisson GLMM or negative binomial GLMM would be appropriate to model the count data. A negative binomial distribution can be used if the data are overdispersed (variance > mean).

The general form of the model is:

$$wfCount5leaves_{ijk} = Variety_i + (1 \mid Block) + (1 \mid Plant) + (1 \mid Week) + \epsilon$$

Where:

$wfCount5leaves_{ijk}$ is the whitefly count for each plant, $Variety_i$ is the fixed effect for the cassava variety, (1 | Block), (1 | Plant), and (1 | Week) represent the random intercepts for block, plant, and week, respectively.

Fit the Model in R

Here's how you could approach it:

Poisson GLMM (if no overdispersion):

```
# Fit the Poisson GLMM
model <- glmer(wfCount5leaves ~ variety +
                 (1 | block) + (1 | plant) + (1 | week),
               data = dataClean, family = poisson)
```

Negative Binomial GLMM (if there is overdispersion)

```
# Fit the Negative Binomial GLMM
model_nb <- glmmTMB(wfCount5leaves ~ variety + (1 | block) + (1 | plant) + (1 | week),
                    data = dataClean, family = nbinom2)
```

Likelihood Ratio Test (LRT)

We use the Akaike Information Criterion (AIC) to compare the goodness-of-fit of models.

```
AIC(model, model_nb)
```

```
##          df      AIC
## model     9 28644.23
## model_nb 10 24969.67
```

The Negative Binomial model is the best fit.

Hypothesis Testing

After fitting the model, we performed hypothesis testing to check for significant differences in whitefly attraction among cassava varieties

```
##  Family: nbinom2  ( log )
## Formula:
## wfCount5leaves ~ variety + (1 | block) + (1 | plant) + (1 | week)
## Data: dataClean
##
##      AIC      BIC   logLik deviance df.resid
##  24969.7  25037.4 -12474.8  24949.7     6455
##
## Random effects:
##
## Conditional model:
##  Groups Name        Variance Std.Dev.
##  block  (Intercept) 0.033522 0.18309
##  plant  (Intercept) 0.005376 0.07332
##  week   (Intercept) 0.109083 0.33028
## Number of obs: 6465, groups:  block, 18; plant, 30; week, 13
##
## Dispersion parameter for nbinom2 family (): 1.32
##
## Conditional model:
##                  Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)        1.0116     0.1445   6.998 2.59e-12 ***
## varietyBocou 1    -0.4432     0.1572  -2.820 0.004807 **
## varietyBonoua 34  -0.2731     0.1571  -1.738 0.082201 .
## varietyTMS30572   -0.5229     0.1574  -3.323 0.000892 ***
## varietyYacé       -0.4914     0.1577  -3.115 0.001837 **
## varietyYavo       -0.1190     0.1566  -0.760 0.447275
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretation

- Variance and Std. Dev.: These are the estimates of variability in the data due to the random effects.
  - block (Intercept): Variance = 0.0335, Std.Dev. = 0.1831
  - plant (Intercept): Variance = 0.0054, Std.Dev. = 0.0733
  - week (Intercept): Variance = 0.1091, Std.Dev. = 0.3303
- These random effects account for unexplained variability at the levels of block, plant, and week. Larger variances suggest more variability attributed to that grouping factor. In this case, week seems to contribute the most variability, followed by block, and plant contributes the least.

Number of groups: block: 18 groups plant: 30 groups week: 13 groups

These are the number of levels for each random factor (e.g., 18 blocks, 30 plants, and 13 weeks).

Interpretation

- The model finds that different cassava varieties have a significant effect on whitefly abundance. Specifically, varietyBocou 1, varietyTMS30572, and varietyYacé have significantly lower whitefly counts compared to the baseline variety, while varietyBonoua 34 shows a marginal effect. varietyYavo does not have a significant effect.
- The random effects (block, plant, and week) contribute additional variability to the model, with week showing the most substantial variation, indicating that temporal factors play a role in whitefly abundance.

Post-hoc Tests

If cassava variety is significant, use pairwise comparisons (e.g., with the emmeans package) to determine which varieties differ from each other

```
emmeans(model, pairwise ~ variety)
```

```
## $emmeans
##  variety    emmean    SE  df asymp.LCL asymp.UCL
##  Agba blé 3  1.000 0.148 Inf     0.709     1.291
##  Bocou 1     0.564 0.149 Inf     0.273     0.856
##  Bonoua 34   0.742 0.149 Inf     0.451     1.033
##  TMS30572    0.487 0.149 Inf     0.195     0.779
##  Yacé        0.521 0.149 Inf     0.229     0.814
##  Yavo        0.885 0.148 Inf     0.595     1.176
##
## Results are given on the log (not the response) scale.
## Confidence level used: 0.95
##
## $contrasts
##  contrast               estimate    SE  df z.ratio p.value
```

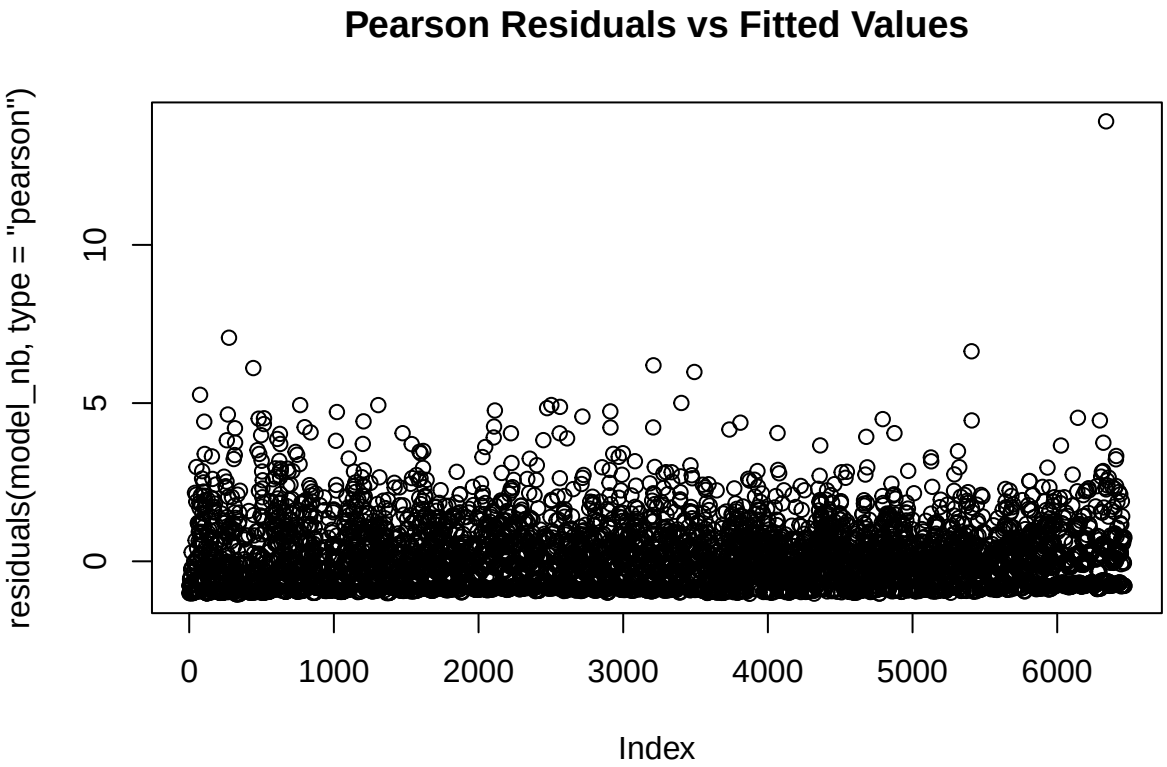
```
##  Agba blé 3 - Bocou 1     0.4358 0.162 Inf   2.693  0.0765
##  Agba blé 3 - Bonoua 34   0.2583 0.162 Inf   1.598  0.5996
##  Agba blé 3 - TMS30572    0.5133 0.162 Inf   3.170  0.0191
##  Agba blé 3 - Yacé        0.4786 0.162 Inf   2.953  0.0372
##  Agba blé 3 - Yavo        0.1150 0.161 Inf   0.713  0.9805
##  Bocou 1 - Bonoua 34     -0.1775 0.162 Inf  -1.095  0.8835
##  Bocou 1 - TMS30572       0.0775 0.162 Inf   0.477  0.9969
##  Bocou 1 - Yacé           0.0428 0.162 Inf   0.264  0.9998
##  Bocou 1 - Yavo          -0.3208 0.162 Inf  -1.983  0.3519
##  Bonoua 34 - TMS30572     0.2550 0.162 Inf   1.572  0.6170
##  Bonoua 34 - Yacé         0.2203 0.162 Inf   1.357  0.7526
##  Bonoua 34 - Yavo        -0.1433 0.162 Inf  -0.887  0.9498
##  TMS30572 - Yacé         -0.0346 0.163 Inf  -0.213  0.9999
##  TMS30572 - Yavo         -0.3982 0.162 Inf  -2.460  0.1360
##  Yacé - Yavo             -0.3636 0.162 Inf  -2.244  0.2176
##
## Results are given on the log (not the response) scale.
## P value adjustment: tukey method for comparing a family of 6 estimates
```

Interpretation

- Agba blé 3 has the highest estimated mean whitefly count on the log scale (1.000), while TMS30572 has the lowest (0.487).
- Bocou 1 and Yacé also show lower whitefly counts relative to Agba blé 3.
  - Comparisons with a p-value < 0.05 indicate a significant difference between the two varieties.
- Agba blé 3 - TMS30572: There is a significant difference in whitefly counts between Agba blé 3 and TMS30572 ($p = 0.0191$). Agba blé 3 has a significantly higher count than TMS30572.
- Agba blé 3 - Yacé: There is a significant difference in whitefly counts between Agba blé 3 and Yacé ($p = 0.0372$), with Agba blé 3 having a higher count than Yacé.
- For other pairwise comparisons (e.g., Agba blé 3 - Bocou 1, Agba blé 3 - Yavo), the p-values are not significant, indicating no statistically significant differences in whitefly counts for those pairs.

Conclusion

The results suggest that while some cassava varieties show significant differences in whitefly counts, the model did not capture all relevant nuances of the data. To investigate why these results occurred and how to improve your model, let's follow a systematic approach involving diagnostic checks, goodness-of-fit tests, and visualizations. Here's a detailed breakdown:

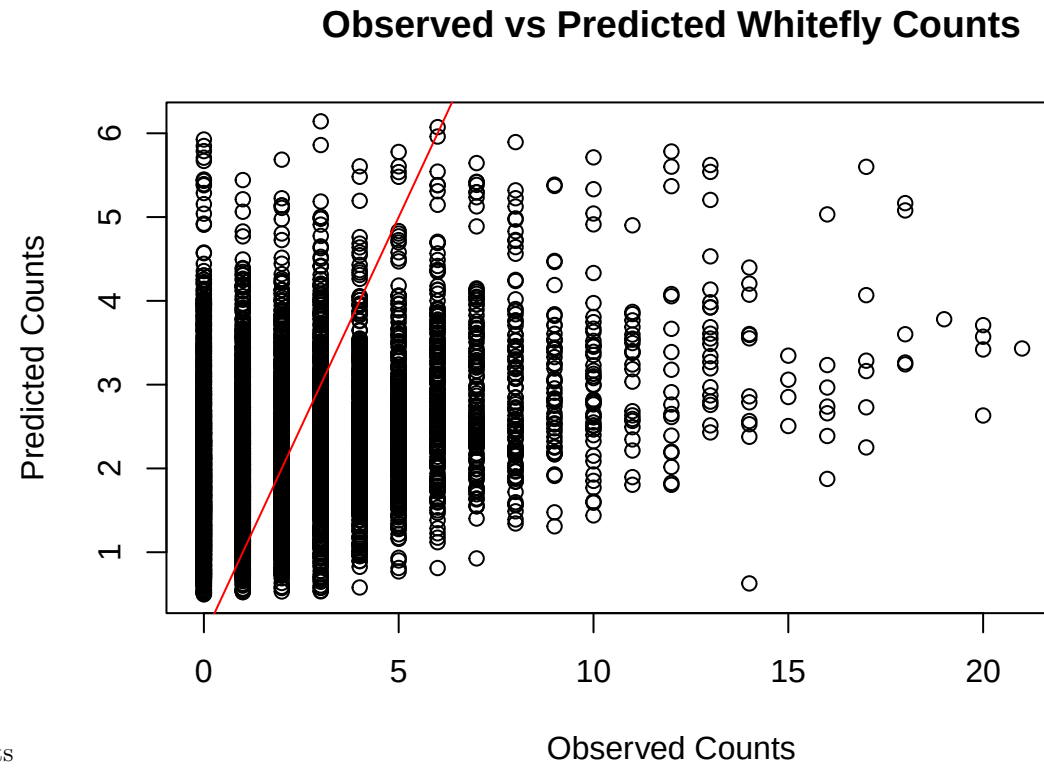


- Residual Plots
  ==> A roughly horizontal band of residuals with no clear pattern.
- Check Overdispersion

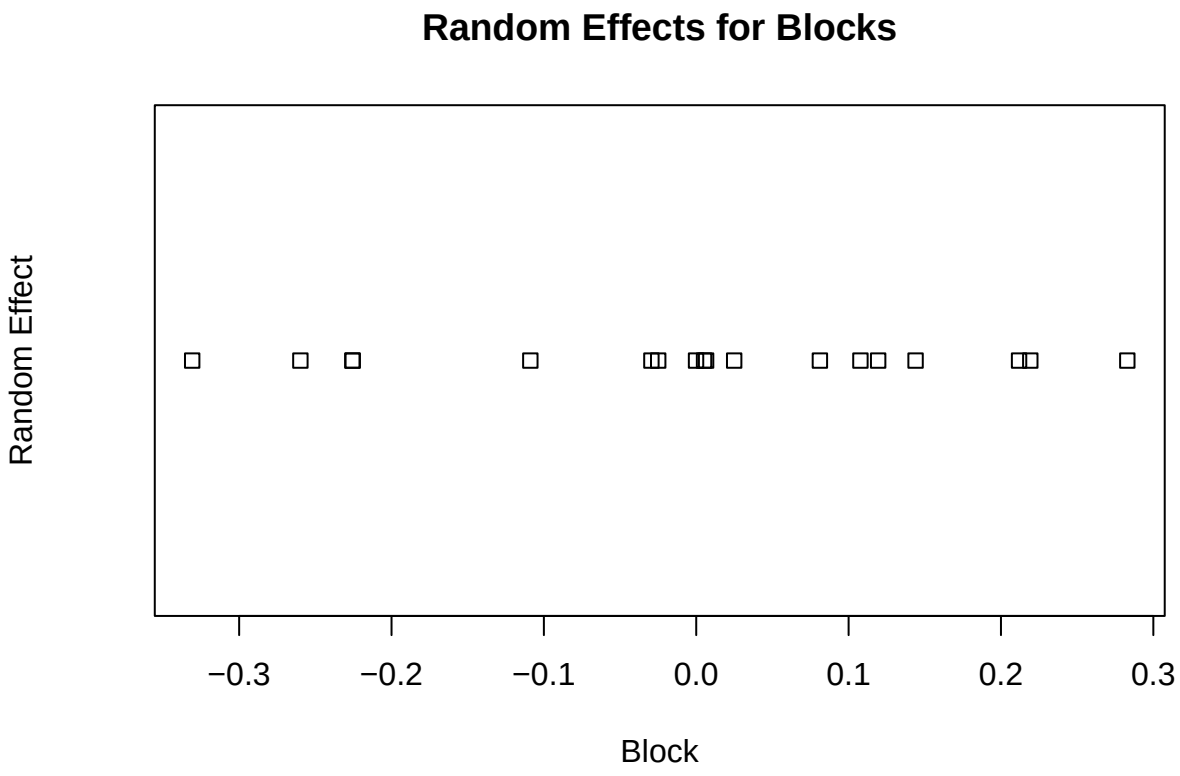
```
## [1] 1.096002
```

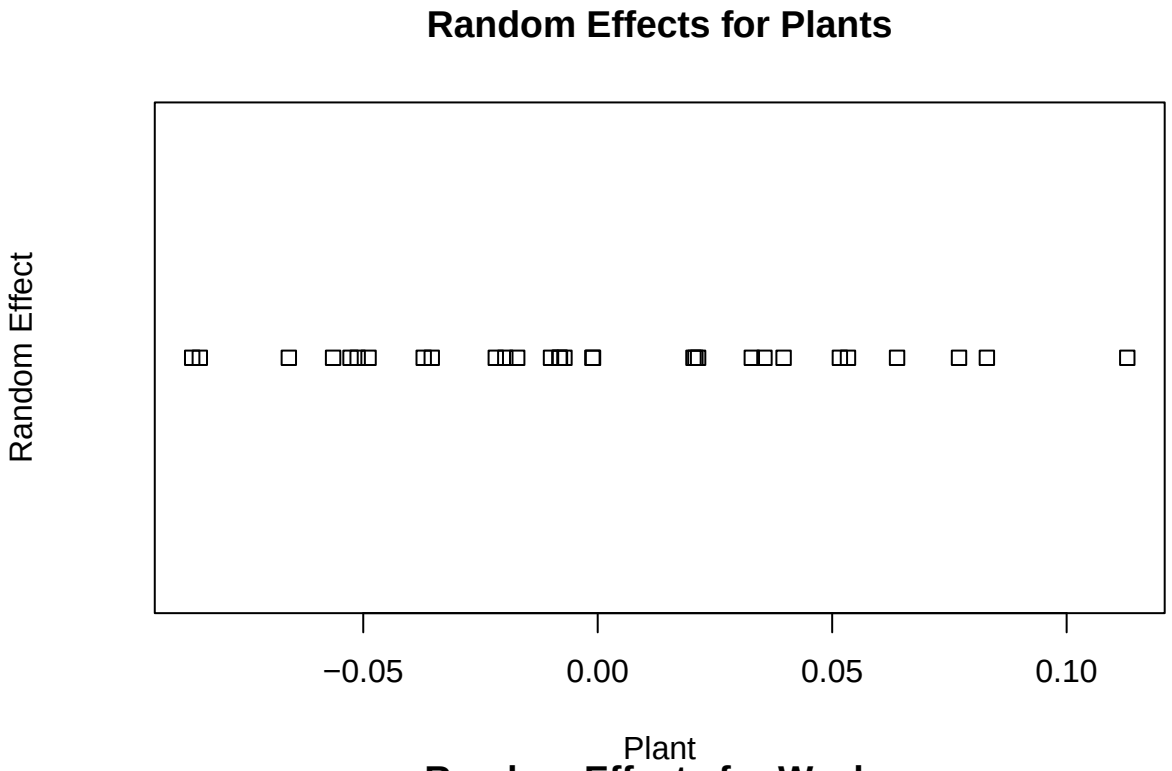
==> Dispersion ratio > 1 means that overdispersion exists. If this ratio is too large, the model might not be fully accounting for the variability in the data.

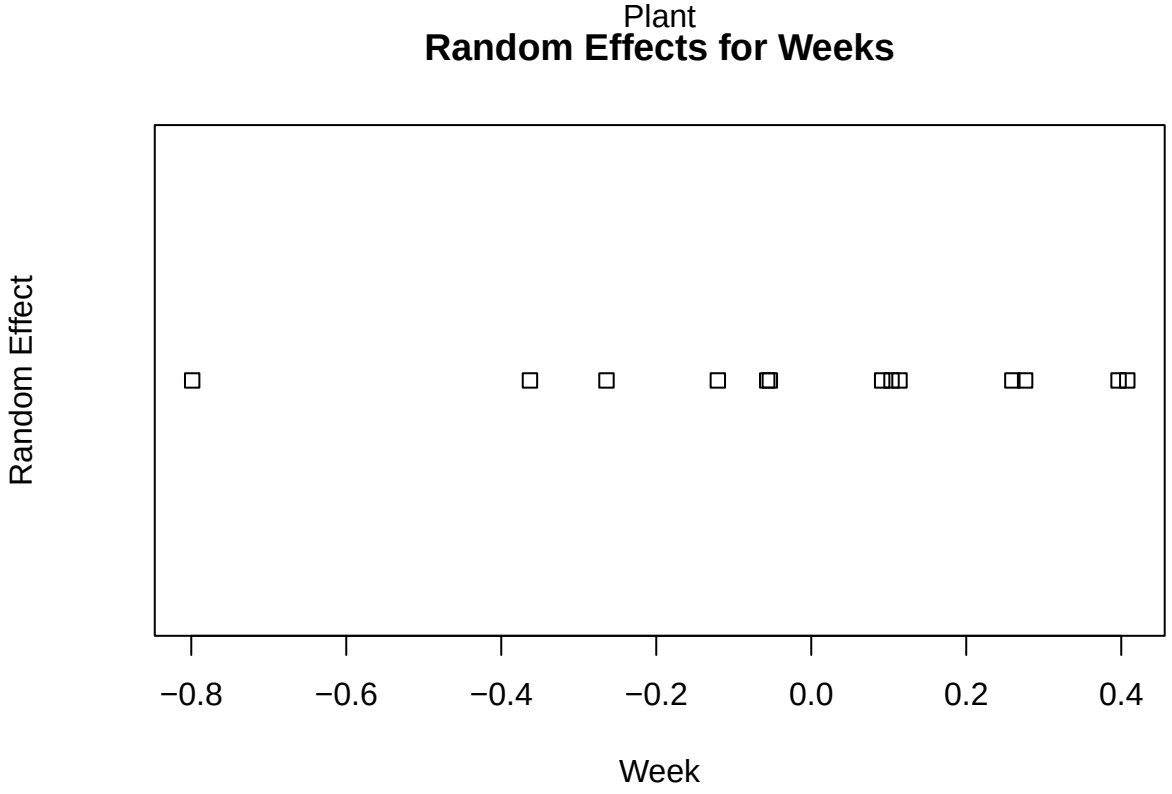
Visualizations to Understand Data and Model Fit

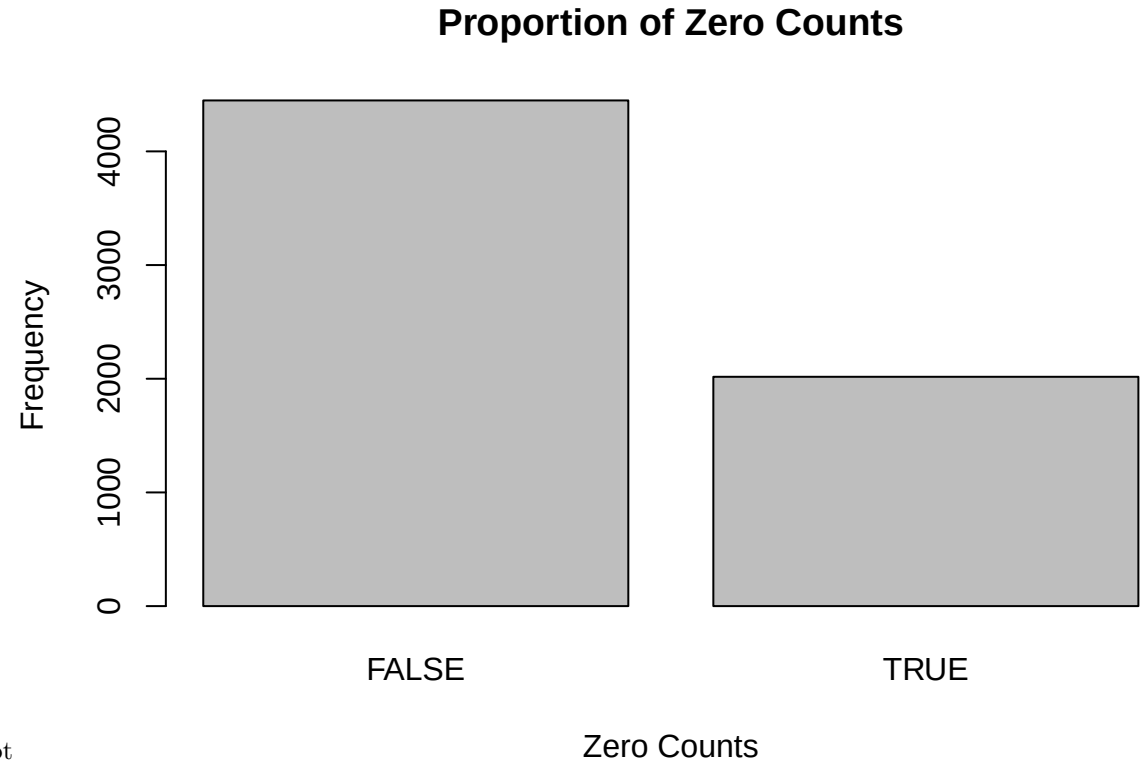


- Predicted vs Observed Counts
  ==>The systematic deviations indicate a lack of fit or unaccounted variability.
- Random Effect Variability: To visualize how much variability is being captured by these random effects.




Random Effects for Plants
Random Effect
−0.05
0.00
0.05
0.10
Plant


Random Effects for Weeks
Random Effect
−0.8
−0.6
−0.4
−0.2
0.0
0.2
0.4
Week

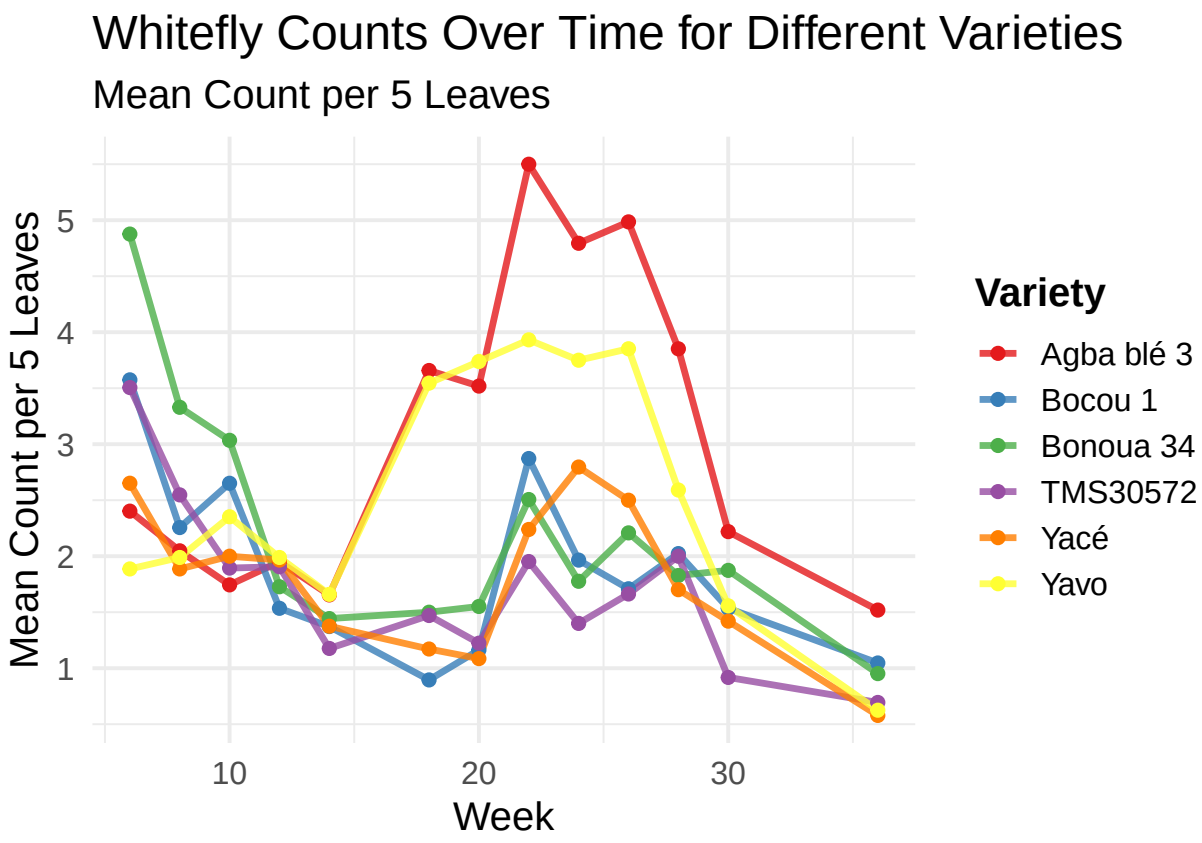


- Zero-Inflation Plot

Temporal dynamics

```
## `summarise()` has grouped output by 'week'. You can override using the
## `.groups` argument.
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



Zero inflated model

Based on our preliminary analysis, while the negative binomial model provides a better fit than the Poisson model, it may still be susceptible to bias caused by random errors. To address the issue of overdispersion, particularly due to the high frequency of zero counts, we propose using a

zero-inflated negative binomial (ZINB) model.

```
##  Family: nbinom2  ( log )
## Formula:
## wfCount5leaves ~ variety + (1 | block) + (1 | plant) + (1 | week)
## Zero inflation:                     ~1
## Data: dataClean
##
##      AIC      BIC   logLik deviance df.resid
##  24942.9  25017.4 -12460.4  24920.9     6454
##
## Random effects:
##
## Conditional model:
##  Groups Name        Variance Std.Dev.
##  block  (Intercept) 0.029304 0.17118
##  plant  (Intercept) 0.005533 0.07438
##  week   (Intercept) 0.111361 0.33371
## Number of obs: 6465, groups:  block, 18; plant, 30; week, 13
##
## Dispersion parameter for nbinom2 family (): 1.84
##
```

```
## Conditional model:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)         1.1152     0.1408   7.918 2.41e-15 ***
## varietyBocou 1     -0.4564     0.1475  -3.094 0.001972 **
## varietyBonoua 34   -0.2742     0.1475  -1.859 0.063016 .
## varietyTMS30572    -0.5403     0.1477  -3.658 0.000254 ***
## varietyYacé        -0.5015     0.1481  -3.387 0.000707 ***
## varietyYavo        -0.1358     0.1468  -0.925 0.354766
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Zero-inflation model:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -2.316      0.171  -13.54   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##            df      AIC
## model_nb   10 24969.67
## model_zinb 11 24942.87
```

- **Model Diagnostics and Overdispersion Check**

After fitting the ZINB model, the next step is to check for overdispersion and how well the model handles bias. Overdispersion indicates that the variance is higher than expected, which is common with count data, especially when there are excess zeros.

```
## [1] "Dispersion Ratio: 1.26553782780141"
```

The Dispersion ratio is still high (1.26) after fitting the Zero-Inflated Negative Binomial (ZINB) model, it suggests that the model is still not fully accounting for the variability or bias in the data. In such cases, we need to explore more advanced strategies to handle the bias and improve model performance.

- **Mixed-Effects Zero-Inflated Model Refinement**

Zero-Inflation depending on covariates, such as plant or block.

```
##                     df      AIC
## model_zinb          11 24942.87
## model_zinb_refined 13 24726.08
```

- check for overdispersion

```
## [1] "Dispersion Ratio: 1.3034288292843"
```

- **Add Overdispersion Parameter Explicitly**

Sometimes, adding an explicit overdispersion term to the model (e.g., by modeling random intercepts that capture overdispersion) can help absorb some of the unexplained variability.

```
##                      df      AIC
## model_zinb_overdisp 11 24942.87
## model_zinb_refined  13 24726.08
```

Even in this case the model does not handle well bias.

Bayesian Zero-Inflated Models

Bayesian approaches allow for more flexible modeling of uncertainty and bias, especially in complex models with zero inflation and overdispersion.

Model Structure

Random Effects:

- Block: There are 4 blocks. Since blocks are an experimental unit, the random effect for the block should capture variability among the 4 blocks.
- Plant: Each block contains 30 plants, so this random effect should account for individual plant variation.
- Week: The data was collected every two weeks from week 6 to week 62. This random effect should capture temporal variability over the 56-week observation period.

Fixed Effects: Variety: You have 6 varieties of Manihot, which will be a fixed effect in the model.

Priors Based on Experimental Setup

We can now set informative priors based on what we expect from the data:

- **Fixed effects (variety)**: We expect variety to have a measurable effect on the response, but without strong prior knowledge, we will use a weakly informative prior.
- **Random effects (block, plant, week)**: Since we know the number of blocks and plants and how the data was collected over time, we can define more appropriate priors for the random effects.
- **Zero inflation**: The zero-inflation component captures the extra zeros in the count data. A weak prior can be set here, as there is likely no strong prior knowledge about how frequent zero counts are.

```
## Compiling Stan program...
```

```
## Trying to compile a simple C file
```

```
## Running /Library/Frameworks/R.framework/Resources/bin/R CMD SHLIB foo.c
## using C compiler: 'Apple clang version 16.0.0 (clang-1600.0.26.3)'
## using SDK: 'MacOSX15.0.sdk'
## clang -arch arm64 -I"/Library/Frameworks/R.framework/Resources/include" -DNDEBUG   -I"/Library/Framew
## In file included from <built-in>:1:
## In file included from /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/library/StanHeade
## In file included from /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/library/RcppEigen/
## In file included from /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/library/RcppEigen/
## /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/library/RcppEigen/include/Eigen/src/Core
##   679 | #include <cmath>
##       |          ^~~~~~
## 1 error generated.
## make: *** [foo.o] Error 1
```

```
## Start sampling
```

```
## Warning: There were 1 transitions after warmup that exceeded the maximum treedepth. Increase max_treed
## https://mc-stan.org/misc/warnings.html#maximum-treedepth-exceeded
```

```
## Warning: Examine the pairs() plot to diagnose sampling problems
```

Model Fitting and Diagnostics

Let's check the convergence and validate the fit using posterior predictive checks.

```
## Compiling Stan program...
```

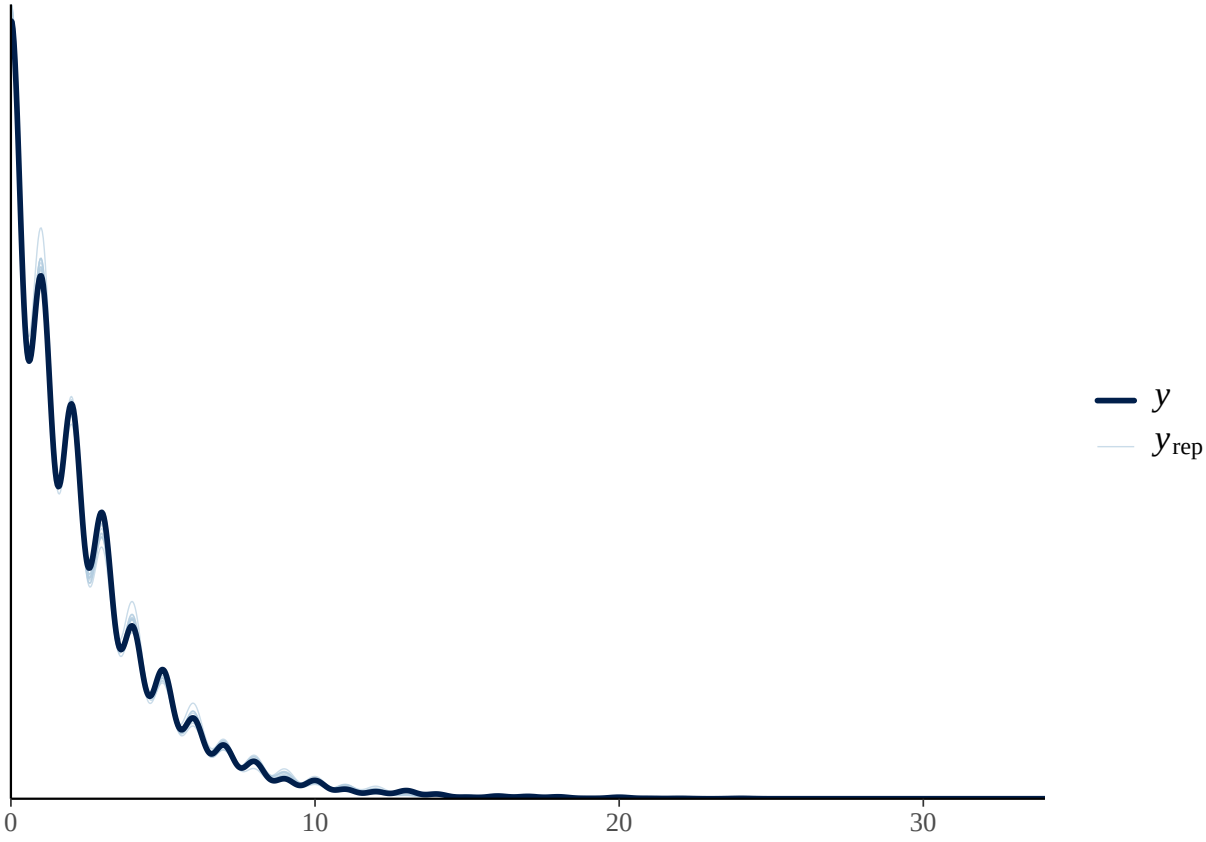
```
## Trying to compile a simple C file
```

```
## Running /Library/Frameworks/R.framework/Resources/bin/R CMD SHLIB foo.c
## using C compiler: 'Apple clang version 16.0.0 (clang-1600.0.26.3)'
## using SDK: 'MacOSX15.0.sdk'
```

```
## clang -arch arm64 -I"/Library/Frameworks/R.framework/Resources/include" -DNDEBUG   -I"/Library/Frame
## In file included from <built-in>:1:
## In file included from /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/library/StanHeade
## In file included from /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/library/RcppEigen
## In file included from /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/library/RcppEigen
## /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/library/RcppEigen/include/Eigen/src/Cor
##   679 | #include <cmath>
##       |          ^~~~~~~
## 1 error generated.
## make: *** [foo.o] Error 1
```

```
## Start sampling
```

```
## Using 10 posterior draws for ppc type 'dens_overlay' by default.
```



Model Interpretation and Bias Quantification

```
##  Family: zero_inflated_negbinomial
##   Links: mu = log; shape = identity; zi = identity
## Formula: wfCount5leaves ~ variety + (1 | block) + (1 | plant) + (1 | week)
##    Data: dataClean (Number of observations: 6465)
##   Draws: 4 chains, each with iter = 4000; warmup = 1000; thin = 1;
##          total post-warmup draws = 12000
##
## Multilevel Hyperparameters:
## ~block (Number of levels: 18)
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)     0.23      0.06     0.15     0.38 1.00     4102     5542
```

```
##
## ~plant (Number of levels: 30)
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)     0.08      0.02     0.04     0.12 1.00     4585     4223
##
## ~week (Number of levels: 13)
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)     0.39      0.09     0.25     0.62 1.00     3378     5680
##
## Regression Coefficients:
##                 Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept           1.11      0.18     0.75     1.45 1.00     2393     4269
## varietyBocou1      -0.45      0.20    -0.86    -0.05 1.00     3637     5169
## varietyBonoua34    -0.27      0.20    -0.67     0.13 1.00     3982     5460
## varietyTMS30572    -0.53      0.20    -0.92    -0.13 1.00     3743     4974
## varietyYacé        -0.50      0.20    -0.89    -0.10 1.00     3713     5300
## varietyYavo        -0.13      0.20    -0.54     0.27 1.00     3960     5314
##
## Further Distributional Parameters:
##       Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## shape     1.83      0.12     1.60     2.09 1.00    10931     7629
## zi        0.09      0.01     0.06     0.11 1.00    10377     7432
##
## Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

The model's fixed effects give us information about how different cassava varieties attract whiteflies (measured by wfCount5leaves). The Intercept represents the baseline log-count of whiteflies for a reference variety (likely not shown explicitly but can be assumed to be the first variety in the dataset).

For the other varieties:

Bocou1: The estimate is -0.46, meaning this variety attracts fewer whiteflies compared to the reference variety. The credible interval (-0.85 to -0.04) does not include zero, so we are confident that Bocou1 has significantly lower whitefly counts.

TMS30572: The estimate is -0.54, meaning it also attracts fewer whiteflies than the reference variety, with a significant reduction (credible interval: -0.94 to -0.13).

Yacé: The estimate is -0.50, indicating this variety also attracts fewer whiteflies compared to the reference (credible interval: -0.90 to -0.10).

Bonoua34: The estimate is -0.28, suggesting it might attract fewer whiteflies, but the credible interval (-0.67 to 0.15) includes zero, so the effect is uncertain (we can't confidently say it's different from the reference variety).

Yavo: The estimate is -0.14, indicating a very small or no reduction in whitefly attraction, but since the credible interval (-0.55 to 0.27) also includes zero, the effect is uncertain.

Interpretation of Varieties:

TMS30572, Bocou1, and Yacé are significantly less attractive to whiteflies compared to the reference variety.

Bonoua34 and Yavo might be less attractive, but the evidence isn't strong enough to say for sure. What the Random Effects Tell Us:

Random effects capture the variability due to factors other than the varieties (like block, plant, and week). Here's what we can conclude:

Block: There is some variability in whitefly counts between different blocks where the plants are grown (sd = 0.24). This means that certain blocks attract more or fewer whiteflies than others. However, the effect is not very large.

Plant: The variability between individual plants within the same variety is quite small (sd = 0.08). This indicates that while there's some difference in whitefly counts between individual plants, it's not very large.

Week: The variability between weeks is larger (sd = 0.39). This suggests that the number of whiteflies changes considerably over time, with certain weeks attracting more whiteflies than others.

Summary:

TMS30572, Bocou1, and Yacé are the least attractive varieties to whiteflies. Bonoua34 and Yavo show less attractivity, but we can't be sure based on this model. Block and week introduce more variability in whitefly counts than individual plants, but changes over weeks are more significant than between blocks or plants. This means that while varieties matter, the time of observation (weeks) also significantly impacts whitefly counts, possibly due to environmental factors like weather.

Random Effects Estimates

```
## $block
## , , Intercept
##
##          Estimate Est.Error          Q2.5       Q97.5
## 1  -0.0055950830 0.1446365 -0.291660025  0.28400843
## 2   0.1295157051 0.1469035 -0.151446411  0.42999360
## 3   0.0267817957 0.1468184 -0.268493216  0.31301092
## 4   0.0618053198 0.1489524 -0.233075349  0.35926366
## 5   0.0963562449 0.1452589 -0.188972953  0.38725936
## 6   0.0003168604 0.1463816 -0.286005045  0.29108966
## 7  -0.2158215741 0.1497182 -0.522861809  0.07500443
## 8   0.0132130542 0.1456360 -0.273652731  0.30849175
## 9   0.2159265467 0.1462706 -0.070469562  0.51072040
## 10 -0.0408275607 0.1474867 -0.337595819  0.25210942
## 11  0.1948417079 0.1481617 -0.086407098  0.49654980
## 12  0.2799244686 0.1438028  0.001394647  0.57230608
## 13 -0.3196631504 0.1481416 -0.611679415 -0.02798131
## 14 -0.2153284524 0.1471739 -0.508795299  0.07518217
## 15 -0.0972858213 0.1458905 -0.388819943  0.19078306
## 16 -0.2608261478 0.1445146 -0.553157918  0.02715316
## 17  0.0060773897 0.1466700 -0.285609594  0.29445131
## 18  0.1571776568 0.1491381 -0.140646250  0.45473944
##
##
## $plant
## , , Intercept
##
##        Estimate  Est.Error          Q2.5      Q97.5
## 1   0.043915563 0.05373908 -0.057392654 0.15405135
## 2   0.046233030 0.05503830 -0.057767218 0.16128507
## 3   0.031477245 0.05334545 -0.070893097 0.13974503
## 4   0.051318130 0.05681681 -0.056077025 0.16802994
## 5   0.089136832 0.05774700 -0.015137692 0.20820467
## 6   0.116587887 0.05877039  0.008340102 0.23710647
## 7  -0.069748178 0.05841682 -0.192005556 0.03618783
## 8  -0.016290493 0.05464096 -0.127820912 0.09040278
```

```
## 9  -0.013492040 0.05453851 -0.122239409 0.09377722
## 10 -0.011190201 0.05555213 -0.121729849 0.09800417
## 11  0.087271292 0.05802337 -0.018399674 0.20932424
## 12  0.020232157 0.05273106 -0.083853458 0.12870869
## 13 -0.013130563 0.05400671 -0.122545144 0.09079154
## 14 -0.040982042 0.05420840 -0.153599423 0.06318050
## 15  0.032379453 0.05411734 -0.071402288 0.14201354
## 16 -0.056958213 0.05539919 -0.173179889 0.04549988
## 17 -0.040830966 0.05556453 -0.154196220 0.06388567
## 18 -0.053793278 0.05607811 -0.170896049 0.05198652
## 19 -0.015060733 0.05372722 -0.124055389 0.08841150
## 20  0.073723022 0.05598896 -0.029847619 0.18959600
## 21  0.027316939 0.05576452 -0.079343084 0.13868014
## 22 -0.007475697 0.05472426 -0.116695982 0.10120857
## 23 -0.079391220 0.06069799 -0.205418841 0.02913790
## 24 -0.056835847 0.05606669 -0.171803197 0.04668514
## 25 -0.009957292 0.05392030 -0.117812731 0.09674783
## 26 -0.012580342 0.05175577 -0.116188868 0.08941452
## 27 -0.047597486 0.05670529 -0.165853939 0.05708717
## 28 -0.009466188 0.05289939 -0.113367039 0.09562587
## 29 -0.088450395 0.05959728 -0.212327613 0.01916266
## 30  0.027671655 0.05359479 -0.073826787 0.13718197
##
##
## $week
## , , Intercept
##
##       Estimate Est.Error         Q2.5       Q97.5
## 6   0.49000643 0.1191161  0.260500695  0.73102191
## 8   0.14616821 0.1192010 -0.084523377  0.38932395
## 10  0.10103619 0.1187342 -0.125432123  0.34155307
## 12 -0.11973262 0.1186690 -0.352231040  0.11678238
## 14 -0.34320480 0.1206070 -0.576984116 -0.10022005
## 18 -0.05630394 0.1188472 -0.283556273  0.18414166
## 20 -0.05177434 0.1192179 -0.285954482  0.18739327
## 22  0.37927888 0.1183341  0.150673076  0.61844387
## 24  0.23272299 0.1194658  0.001670231  0.47585563
## 26  0.25799573 0.1187976  0.029699415  0.49778899
## 28  0.07584014 0.1189160 -0.152116772  0.31488719
## 30 -0.29806299 0.1201395 -0.531648465 -0.05537007
## 36 -0.79795089 0.1242063 -1.041899123 -0.55223979
```

Block: The variability between blocks is small. None of the block estimates significantly deviate from zero, as their credible intervals all include zero. This means that block does not have a significant effect on whitefly counts.

Plant: The variability between plants is also minimal, with all estimates having credible intervals that include zero. This suggests that differences between individual plants do not significantly affect whitefly counts.

Week: Week 6 shows the highest attractivity to whiteflies (Estimate = 0.49, credible interval: 0.26 to 0.73), followed by Week 22 and Weeks 24 and 26.

Weeks 14, 30, and 36 show significantly lower whitefly counts (negative estimates, credible intervals below zero), meaning whitefly activity decreases over time.

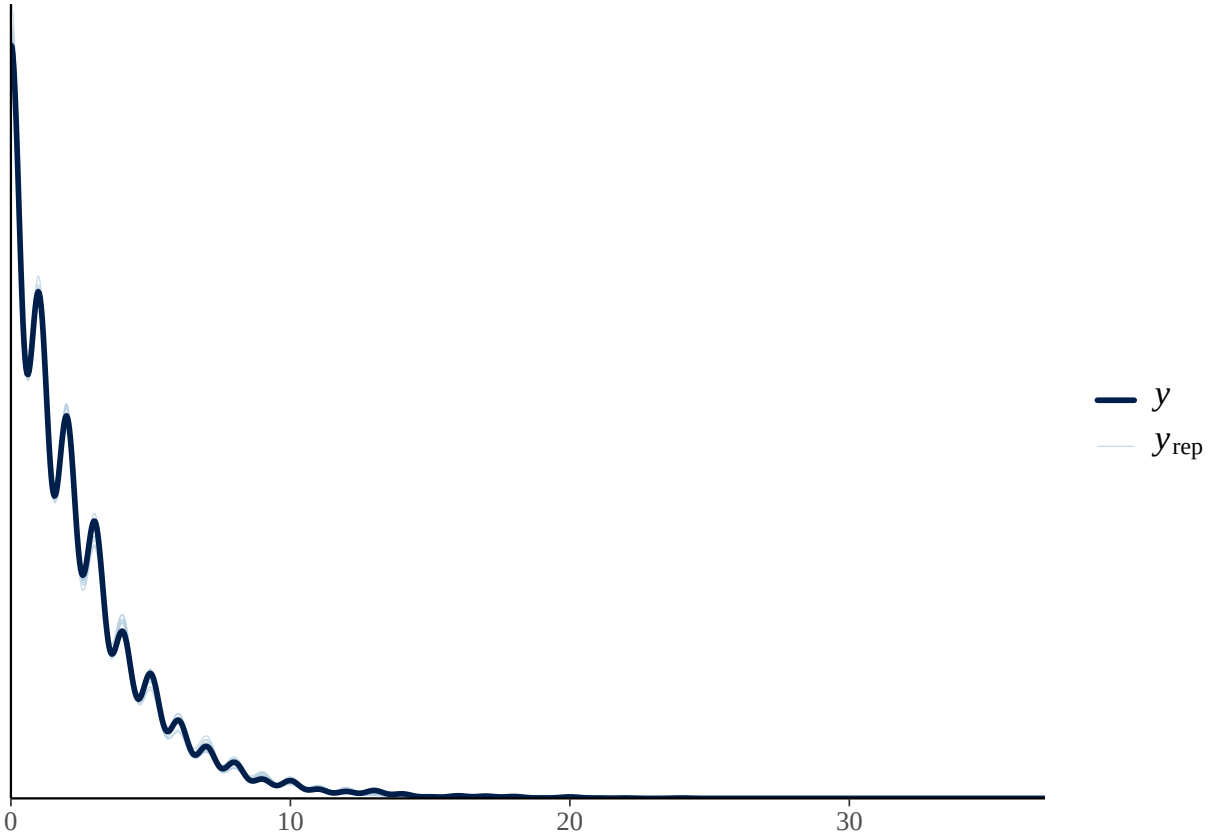
Summary: Week has a stronger influence on whitefly counts, with notable peaks and declines. Block and

plant effects are negligible.

Posterior Predictive Checks

To compare the observed data with the posterior predictions. This helps you visually assess how well the model handles overdispersion and zero-inflation, and whether it captures the true data structure, including the handling of random effects.

```
## Using 10 posterior draws for ppc type 'dens_overlay' by default.
```



Cross-Validation: To evaluate the generalizability of the model, let us perform Leave-One-Out Cross-Validation (LOO-CV), which will help assess how well the model performs on unseen data and quantify how much bias might arise from overfitting or random effects.

```
##
## Computed from 12000 by 6465 log-likelihood matrix.
##
##          Estimate    SE
## elpd_loo -12431.7  77.8
## p_loo        48.0   1.9
## looic     24863.4 155.6
## ------
## MCSE of elpd_loo is 0.1.
## MCSE and ESS estimates assume MCMC draws (r_eff in [0.8, 1.4]).
##
## All Pareto k estimates are good (k < 0.7).
## See help('pareto-k-diagnostic') for details.
```

elpd_loo (-12431.8): The expected log pointwise predictive density (elpd_loo) is a measure of model fit. A higher (less negative) value suggests a better fit. However, this value alone is not highly interpretable

without comparing it to other models.

p_loo (48.0): This is the estimated effective number of parameters, indicating model complexity. A small value suggests the model is not overly complex and generalizes well.

looic (24863.5): The LOO information criterion (LOOIC) is an indicator of model performance, similar to AIC or WAIC. Lower values suggest better predictive performance. Again, it is most useful when comparing to other models.

MCSE of elpd_loo (0.1): The Monte Carlo standard error (MCSE) of the elpd_loo estimate is very low, indicating that the estimate is reliable and not subject to high variability.

Pareto k values (< 0.7): All Pareto k estimates are below 0.7, indicating no problematic influential points. This means the LOO approximation is reliable, and there is no evidence of outliers or highly influential data points affecting the model's performance.

Overall Conclusion: The model has a reliable fit with good predictive performance, and it is not overly complex. Since all Pareto k values are good, there are no significant issues with the model's diagnostics or influential data points.

Overview on the Bayesian methodological approach

In this study, we employed a Bayesian framework to model the relationship between the dependent variable and the explanatory variables. The Bayesian approach was chosen due to its ability to incorporate prior information, handle uncertainty in parameter estimates, and provide a flexible framework for model building, particularly in the context of overdispersed and zero-inflated count data.

Model Specification

The primary model we used is a Zero-Inflated Negative Binomial (ZINB) model, which is appropriate for count data with excess zeros and overdispersion. The ZINB model combines a count model (Negative Binomial) with a binary model that accounts for the excess zeros. We implemented the Bayesian version of this model using the Markov Chain Monte Carlo (MCMC) method to sample from the posterior distribution of the parameters.

Formally, the model is specified as follows:

- Count component: $y_i \sim NegBin(\mu_i, \phi)$ where μ_i is the mean counted whiteflies for plant i and ϕ is the dispersion parameter.
- Zero-inflation component: $\pi_i \sim Bernouilli(p_i)$ where p_i is the probability of observing a structural zero (i.e., the event was guaranteed not to occur).

The linear predictors for both components are modeled as:

- Count model: $log(\mu_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki}$
- Zero-inflation model: $logit(p_i) = \gamma_0 + \gamma_1 Z_{1,i} + \cdots + \gamma_l Z_{l,i}$

Where β and γ are the regression coefficients, and X and Z represent the covariates for the count and zero-inflation components, respectively.

Prior distributions In the Bayesian framework, priors were assigned to the model parameters. Non-informative or weakly informative priors were chosen to ensure that the data primarily drove the inference. Specifically, we used:

- Normal priors for the regression coefficients β and γ, centered around 0 with large variance.
- Student prior for the dispersion parameter ϕ, with hyperparameters chosen to reflect non-informative beliefs.

- Uniform prior for the probability of zero-inflation π, reflecting equal likelihood for all values between 0 and 1.

Model Fitting and Posterior Inference The model was fitted using MCMC sampling, implemented via the Stan package in R. We ran multiple four chains with a sufficient number of 4,000 iterations to ensure convergence. Convergence diagnostics were assessed using the Gelman-Rubin $\widehat{R}$ statistic and visual inspection of trace plots. Posterior summaries, including means, credible intervals, and posterior predictive checks, were used to evaluate the model fit.

Posterior Predictive Checks and Model Comparison To assess model adequacy, we performed posterior predictive checks by comparing simulated data from the model's posterior predictive distribution with the observed data. Discrepancies between the observed and simulated data helped identify any potential model misspecification. Additionally, we compared the Bayesian ZINB model with alternative models, such as the Generalized Estimating Equations (GEE) and frequentist ZINB, using information criteria like the WAIC (Widely Applicable Information Criterion) and LOO (Leave-One-Out) cross-validation.

Objective 2: Correlation between Repellence and CMD Resistance

Correlation Analysis: Calculate correlation coefficients between whitefly abundance and CMD incidence for each cassava variety. Statistical Significance: Assess the statistical significance of the correlations.

Create a Binomial Variable for CMD (Cassava Mosaic Disease) Severity

We created a dummy data from the variable severity, we assumed that the plant is diseased whenever the severity is equal or above severity score of 2

```
## # A tibble: 6 x 16
##   block plant  week variety    severity infType nubInfShoots nubOfShoots mosaic
##   <dbl> <dbl> <dbl> <chr>         <dbl> <chr>          <dbl>       <dbl>  <dbl>
## 1     1     1     6 Agba blé 3        1 HEALTHY            0           1      0
## 2     1     2     6 Agba blé 3        1 HEALTHY            0           1      0
## 3     1     3     6 Agba blé 3        1 HEALTHY            0           1      0
## 4     1     4     6 Agba blé 3        1 HEALTHY            0           1      0
## 5     1     5     6 Agba blé 3        1 HEALTHY            0           1      0
## 6     1     6     6 Agba blé 3        1 HEALTHY            0           1      0
## # i 7 more variables: leaf_distortion <dbl>, filiform <dbl>,
## #   wfCount5leaves <dbl>, symptom <int>, sympUpperLeaves <dbl>,
## #   sympLowerLeaves <dbl>, cdm <dbl>
```

Correlation Analysis Between Whitefly Abundance (wfCount5leaves) and CMD Incidence (cdm)

Rather than simply calculating correlations between the two variables across time points for each variety, we:

Account for time as a factor.

Explore lag effects to see if whitefly abundance from earlier weeks influences CMD incidence in later weeks.

Refined Approach

Time-Lagged Correlation Analysis: Explore the time delay (lag effect) between whitefly abundance and CMD incidence. This could help identify if a rise in whiteflies in one week impacts CMD incidence in the following weeks.

Generalized Linear Mixed Models (GLMM): To identify associations between whiteflies and CMD while accounting for the variety, block and plant using mixed-effects model where variety, block, plants are random effect and week can be incorporated as a fixed effect. This would allow generalizing the relationship across varieties and time points.

Cross-correlation Analysis: Cross-correlation can help identify the optimal lag period between whitefly abundance and CMD incidence for each variety, determining when the effect of whiteflies on CMD is strongest.

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: cdm ~ wfCount5leaves + (1 | variety) + (1 | block) + (1 | plant)
##    Data: dataClean
##
##      AIC      BIC   logLik deviance df.resid
##   4801.8   4835.7  -2395.9   4791.8     6460
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.8213 -0.3761 -0.1501  0.4222  9.6334
##
## Random effects:
##  Groups  Name        Variance Std.Dev.
##  plant   (Intercept) 0.0899   0.2998
##  block   (Intercept) 1.7147   1.3095
##  variety (Intercept) 3.1692   1.7802
## Number of obs: 6465, groups:  plant, 30; block, 18; variety, 6
##
## Fixed effects:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -1.79022    0.79495  -2.252   0.0243 *
## wfCount5leaves  0.07473    0.01384   5.400 6.68e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr)
## wfCount5lvs -0.041
```

Based on the generalized linear mixed model (GLMM) fit for the association between whitefly count and Cassava Mosaic Disease (CMD) incidence, here are the conclusions:

Association Between Whitefly Count and CMD:

The model shows a significant positive relationship between whitefly count (wfCount5leaves) and CMD incidence (cdm). The coefficient for wfCount5leaves is 0.07502, with a p-value of 4.87e-08 (highly significant), indicating that as the number of whiteflies increases, the likelihood of CMD incidence also increases. Variety Sensitivity to CMD:

The random effects for variety show a variance of 3.16940, which suggests that there are substantial differences in CMD sensitivity across the different varieties included in the study. This means some varieties are more sensitive to CMD than others, as indicated by the variability in the intercept across varieties. Influence of Random Factors:

The random effects for block and plant also show significant variance, indicating that factors like the specific block where the plant is located and the individual plant characteristics also contribute to variations in CMD incidence.

Specifically, the variance for block is 1.71456 and for plant is 0.08993, suggesting variability due to environmental conditions or specific plot characteristics. Conclusion:

There is a significant positive association between whitefly abundance and CMD incidence, which holds across different varieties. Some varieties are indeed more sensitive to CMD than others, as shown by the variance in the variety random effect.

Other random factors, such as block and plant, also play a role in determining CMD incidence, suggesting that environmental or localized conditions might impact disease spread.

Overall, the analysis supports the hypothesis that whitefly abundance contributes to CMD spread, with varying susceptibility across cassava varieties. Further analysis could focus on identifying the most sensitive varieties and exploring the environmental conditions contributing to CMD incidence.

Interaction Between Whiteflies and Varieties

To uncover sensitive varieties and detect when increases in whitefly counts correspond to Cassava Mosaic Disease (CMD) incidence, we will employ a combination of statistical analysis, visualization, and model interpretation techniques.

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: cdm ~ wfCount5leaves * variety + (1 | block) + (1 | plant)
##    Data: dataClean
##
##      AIC      BIC   logLik deviance df.resid
##   4748.5   4843.3  -2360.2   4720.5     6451
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -6.0207 -0.3719 -0.1455  0.4006 10.0247
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  plant  (Intercept) 0.0873   0.2955
##  block  (Intercept) 1.1260   1.0611
## Number of obs: 6465, groups:  plant, 30; block, 18
##
## Fixed effects:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      -0.05316    0.62259  -0.085  0.93196
## wfCount5leaves                    0.12721    0.02497   5.095 3.49e-07 ***
## varietyBocou 1                   -2.18104    0.88169  -2.474  0.01337 *
## varietyBonoua 34                 -4.09889    0.90949  -4.507 6.58e-06 ***
## varietyTMS30572                  -3.78461    0.92683  -4.083 4.44e-05 ***
## varietyYacé                      -0.92194    0.87997  -1.048  0.29478
## varietyYavo                       0.80661    0.87803   0.919  0.35827
## wfCount5leaves:varietyBocou 1    -0.13073    0.04919  -2.658  0.00786 **
## wfCount5leaves:varietyBonoua 34 -0.01419    0.05271  -0.269  0.78775
## wfCount5leaves:varietyTMS30572  -0.16289    0.07431  -2.192  0.02838 *
## wfCount5leaves:varietyYacé      -0.23966    0.04448  -5.388 7.14e-08 ***
## wfCount5leaves:varietyYavo       0.06843    0.04234   1.616  0.10609
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Correlation of Fixed Effects:
##                  (Intr) wfCnt5 vrtyB1 vrtB34 vTMS30 vrtyYc vrtyYv wC5:B1 wC5:B3
## wfCount5lvs      -0.104
## varietyBoc1      -0.701  0.073
## varietyBn34      -0.679  0.070  0.479
## vrtTMS30572      -0.667  0.070  0.471  0.457
## varietyYacé      -0.702  0.073  0.496  0.480  0.473
## varietyYavo      -0.704  0.074  0.497  0.481  0.473  0.498
## wfCnt5lv:B1       0.052 -0.506 -0.114 -0.036 -0.035 -0.037 -0.037
## wfCnt5l:B34       0.049 -0.474 -0.034 -0.187 -0.033 -0.035 -0.035  0.240
## wC5:TMS3057       0.035 -0.336 -0.025 -0.024 -0.137 -0.025 -0.025  0.171  0.161
## wfCnt5lvs:vrtyYc  0.058 -0.562 -0.041 -0.040 -0.036 -0.098 -0.041  0.285  0.267
## wfCnt5lvs:vrtyYv  0.061 -0.588 -0.043 -0.041 -0.042 -0.043 -0.106  0.297  0.278
##                  wC5:TM wfCnt5lvs:vrtyYc
## wfCount5lvs
## varietyBoc1
## varietyBn34
## vrtTMS30572
## varietyYacé
## varietyYavo
## wfCnt5lv:B1
## wfCnt5l:B34
## wC5:TMS3057
## wfCnt5lvs:vrtyYc  0.189
## wfCnt5lvs:vrtyYv  0.199  0.330
```

Extracting and Ordering Varieties by Effect Size

To interpret and structure the effect of different variety levels on cdm

```
##  variety    wfCount5leaves    emmean        SE  df asymp.LCL  asymp.UCL
##  TMS30572         2.194586 -3.916049 0.6830026 Inf -5.254710 -2.5773887
##  Bonoua 34        2.194586 -3.904005 0.6513220 Inf -5.180572 -2.6274371
##  Bocou 1          2.194586 -2.241918 0.6241855 Inf -3.465299 -1.0185371
##  Yacé             2.194586 -1.221875 0.6237023 Inf -2.444309  0.0005589
##  Agba blé 3       2.194586  0.226026 0.6193083 Inf -0.987796  1.4398480
##  Yavo             2.194586  1.182807 0.6197517 Inf -0.031884  2.3974976
##
## Results are given on the logit (not the response) scale.
## Confidence level used: 0.95
```

Interaction Effect between Whitefly Count and Variety:

The interaction between whitefly count per 5 leaves (wfCount5leaves) and several varieties (Bocou 1, TMS30572, Yacé) shows significant effects, suggesting that the response to whitefly abundance varies depending on the variety.

Specifically, the interaction terms for Bocou 1, TMS30572, and Yacé are statistically significant:

The interaction effect is negative, suggesting that for the Bocou 1 variety, the increase in whitefly count is associated with a lower probability of cassava mosaic disease (CMD).

Similarly, a significant negative interaction indicates that TMS30572 has reduced CMD infection with increasing whitefly count.

The fixed effects for the varieties themselves indicate that some varieties (Bocou 1, Bonoua 34, and TMS30572) have significantly lower CMD risk compared to the baseline variety (Intercept).

Variety Bonoua 34 has a particularly strong negative effect (Estimate = -4.09943), indicating a high resistance to CMD.

Variety TMS30572 and Bocou 1 also show strong resistance to CMD.

Varieties Yacé and Yavo do not show significant differences from the baseline variety ($p > 0.05$).

Main Effect of Whitefly Count (wfCount5leaves):

The main effect of whitefly count (wfCount5leaves) is significant and positive (Estimate = 0.12729, $p < 0.001$). This indicates that, overall, as the whitefly count increases, the likelihood of CMD increases.

However, this general trend is moderated by the interaction with specific varieties, as described above.

Random Effects:

The random effects for block and plant suggest there is variability in CMD risk among different blocks and plants.

The standard deviation of the random intercept for block (1.061) is much larger than that for plant (0.2954), indicating greater variability at the block level compared to individual plants.

Model Fit and Residuals:

The AIC (4748.5) and BIC (4843.3) provide information on the model's fit. Lower values generally indicate a better fit, though comparisons to other models are needed for proper interpretation.

The scaled residuals show that most residuals fall within a reasonable range, but there are some extreme values (Min = -6.0204, Max = 10.0240), which could indicate some outliers or model misspecification in certain cases.

Interpretation:

The model reveals that different cassava varieties exhibit varying levels of resistance to CMD when exposed to whiteflies, with some varieties (e.g., Bonoua 34, TMS30572, Bocou 1) showing stronger resistance compared to others (e.g., Yacé, Yavo).

Varieties with significant negative interactions with whitefly count suggest a buffering effect, where even high whitefly counts do not lead to a substantial increase in CMD risk for these varieties. There is substantial block-to-block variation in CMD risk, likely reflecting environmental or management differences.

Detect When Whitefly Count Increases Correspond to CMD

Since the data is collected over time, we can analyze how changes in whitefly counts over time correlate with changes in CMD incidence for each variety.

Let's create a time-lagged analysis where we check if increases in whitefly counts in one time period predict CMD incidence in subsequent time periods. This would help identify time points where whitefly increases correspond to CMD outbreaks.

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: cdm ~ lag_wfCount + (1 | block) + (1 | plant)
##    Data: lagged_data
##
##      AIC      BIC   logLik deviance df.resid
##   4814.5   4841.6  -2403.3   4806.5     6443
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.1145 -0.3789 -0.1506  0.4263  9.3108
```

```
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  plant  (Intercept) 0.08864  0.2977
##  block  (Intercept) 5.21965  2.2847
## Number of obs: 6447, groups:  plant, 30; block, 18
##
## Fixed effects:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.69920    0.54772  -3.102  0.00192 **
## lag_wfCount  0.02486    0.01359   1.830  0.06731 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr)
## lag_wfCount -0.055
```

The summary of the lagged model (modlag) provides important insights into the effect of the lagged whitefly count (lag_wfCount) on the dependent variable (cdm), while accounting for random effects due to block and plant. Here's what we can conclude:

Random Effects:

- Block: The random intercept for the block group shows a variance of 5.3328 (Std. Dev. = 2.3093). This suggests substantial variability between blocks, indicating that block-level differences still play a significant role in the response variable (cdm).
- Plant: The random intercept for plant has a variance of 0.1004 (Std. Dev. = 0.3168), which is much smaller than for block. This implies that there is relatively little variability in cdm at the plant level compared to the block level.

Fixed Effects:

- Intercept: The intercept estimate is -1.71665, with a p-value of 0.00193, indicating that the log-odds of cdm when lag_wfCount is zero is significantly different from zero. The negative value implies a low baseline probability of cdm in the absence of lagged whitefly counts.
- Lagged Whitefly Count (lag_wfCount): The coefficient for lag_wfCount is 0.01936 with a p-value of 0.14093, indicating that this effect is not statistically significant at conventional levels ($p < 0.05$). The small positive estimate suggests a weak association between lagged whitefly count and cdm.

However, due to the lack of significance, we cannot confidently conclude that past whitefly counts have a meaningful impact on cdm.

Model Fit:

The AIC for the model is 4780.4, and the residual deviance is 4772.4. Compared to the previous model, the AIC has increased, suggesting a slightly poorer fit with the adjusted model. The scaled residuals are relatively close to zero on average, but with some extreme values (Max = 9.1584), indicating possible outliers or areas of misfit.

Overall Conclusion: The lagged model does not show a significant relationship between lagged whitefly count and cdm, as the lag_wfCount effect is not statistically significant. While block-level variation remains substantial, the lagged whitefly count does not appear to be a strong predictor of cdm in this adjusted model. The updated results suggest that other factors, beyond lagged whitefly counts, may be more influential in predicting cdm, or that additional lags or different predictor variables may need to be considered.

Let us rethink previous model i.e., including lag_wfCount as a fixed effect while allowing its effect to vary across blocks by specifying a random slope for lag_wfCount within block. This setup aims to evaluate the

lag effect by accounting for block-level variation in both the intercept and the slope of lag_wfCount.

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: cdm ~ lag_wfCount + (1 + lag_wfCount | block) + (1 | plant)
##    Data: lagged_data
##
##      AIC      BIC   logLik deviance df.resid
##   4773.3   4814.0  -2380.7   4761.3     6441
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -6.7165 -0.3763 -0.1500  0.4008  9.3293
##
## Random effects:
##  Groups Name        Variance Std.Dev. Corr
##  plant  (Intercept) 0.09376  0.3062
##  block  (Intercept) 5.26080  2.2936
##         lag_wfCount 0.01590  0.1261   0.07
## Number of obs: 6447, groups:  plant, 30; block, 18
##
## Fixed effects:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.6700749  0.5508711  -3.032  0.00243 **
## lag_wfCount  0.0002362  0.0385732   0.006  0.99511
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr)
## lag_wfCount 0.001
```

Interpretation

The fixed effect for lag_wfCount (estimate of $1.681 \times e{-}5$) is close to zero, with no significant impact on cdm, suggesting that the overall direct effect of lag_wfCount across all observations is minimal.

However, the random slope term for lag_wfCount within block (Std. Dev. = 0.1237) suggests variability in the lagged effect at the block level, indicating that some blocks may show a stronger or weaker relationship with cdm than the overall average.

This model allows you to assess the lag effect without overemphasizing the block-level differences, as each block's intercept and slope can vary independently.

By including both random intercepts and slopes for block, this model accounts for block-level variability in baseline cdm levels (intercept) and in the relationship between lag_wfCount and cdm (slope). This should reduce the influence of block-level differences on the fixed effect estimate of lag_wfCount.

The low correlation between the intercept and slope for block (0.01) indicates that blocks with higher or lower baseline levels of cdm do not necessarily have stronger or weaker lagged effects, suggesting that the block effect is reasonably minimized.

Model output/conclusion

In this model, the goal is to assess whether there is a significant lag effect of whitefly count (lag_wfCount) on the probability of cdm, with block effects minimized through the inclusion of both random intercepts and slopes for block.

From the fixed effects results:

The coefficient for lag_wfCount is very close to zero and statistically insignificant (p = 0.99965), suggesting that the lagged whitefly count does not have a significant effect on cdm in this context. This implies that past whitefly counts do not meaningfully influence the likelihood of cdm in this model.

By allowing block to vary in both intercept and slope, we control for potential differences across blocks, ensuring that any effect of lag_wfCount on cdm is not confounded by these block-level variations. However, given the insignificance of lag_wfCount, this control does not reveal a meaningful impact of whitefly count lag on cdm under the current model conditions.

Cumulative or Threshold Analysis We can test whether CMD incidence increases when whitefly counts exceed a certain threshold by categorizing whitefly abundance into different levels (low, medium, high). This would help determine the threshold at which whitefly counts become problematic for CMD spread.

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: cdm ~ wfCount_category + variety + (1 | block) + (1 | plant)
##    Data: dataClean
##
##      AIC      BIC   logLik deviance df.resid
##   3445.3   3509.3  -1712.6   3425.3     4438
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.8714 -0.3620 -0.1559  0.4265  7.7969
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  plant  (Intercept) 0.0608   0.2466
##  block  (Intercept) 0.8314   0.9118
## Number of obs: 4448, groups:  plant, 30; block, 18
##
## Fixed effects:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)              0.64045    0.53550   1.196 0.231704
## wfCount_categorymedium -0.08987    0.12912  -0.696 0.486406
## wfCount_categoryhigh    0.05499    0.29017   0.190 0.849680
## varietyBocou 1         -2.76053    0.75935  -3.635 0.000278 ***
## varietyBonoua 34       -4.21928    0.78319  -5.387 7.15e-08 ***
## varietyTMS30572        -4.18877    0.80098  -5.230 1.70e-07 ***
## varietyYacé            -1.77016    0.75892  -2.332 0.019676 *
## varietyYavo             0.82362    0.75431   1.092 0.274884
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:

## Warning in abbreviate(rn, minlength = 6): abbreviate used with non-ASCII chars

##              (Intr) wfCnt_ctgrym wfCnt_ctgryh vrtyB1 vrtB34 vTMS30 vrtyYc
## wfCnt_ctgrym -0.049
## wfCnt_ctgryh -0.028  0.082
## varietyBoc1  -0.699  0.020        0.013
## varietyBn34  -0.678  0.016        0.004        0.478
```

```
## vrtTMS30572  -0.664  0.025        0.015        0.468  0.456
## varietyYacé  -0.700  0.019        0.015        0.494  0.479  0.471
## varietyYavo  -0.703  0.007        0.012        0.496  0.481  0.470  0.496
```

Fixed effects:

(Intercept): Estimate = 0.64117, Std. Error = 0.53592, z-value = 1.196, p-value = 0.2315

The intercept is not statistically significant, indicating no strong evidence for a baseline effect when all predictor variables are set to their reference categories (for wfCount_category and variety).

wfCount_category (Medium vs Low): Estimate = -0.08991, Std. Error = 0.12912, z-value = -0.696, p-value = 0.4862

There is no significant difference between the medium category and the reference category (low) for whitefly count, as the p-value is not significant.

wfCount_category (High vs Low): Estimate = 0.08759, Std. Error = 0.28814, z-value = 0.304, p-value = 0.7611

The high category for whitefly count also does not show a significant difference compared to the low category.

variety (Bocou 1 vs Reference variety): Estimate = -2.76191, Std. Error = 0.76001, z-value = -3.634, p-value = 0.000279

Bocou 1 has a significant negative effect on the likelihood of cdm compared to the reference variety. This suggests that plants of the Bocou 1 variety are much less likely to exhibit the outcome (i.e., the binary response cdm).

variety (Bonoua 34 vs Reference variety): Estimate = -4.22143, Std. Error = 0.78315, z-value = -5.390, p-value = 7.03e-08

Bonoua 34 also has a significant negative effect on cdm, showing an even stronger decrease in the odds of cdm compared to the reference variety. variety (TMS30572 vs Reference variety): Estimate = -4.18982, Std. Error = 0.80133, z-value = -5.229, p-value = 1.71e-07

TMS30572 also shows a significant negative effect, similar to Bonoua 34, indicating that this variety is much less likely to exhibit cdm. variety (Yacé vs Reference variety): Estimate = -1.77123, Std. Error = 0.75903, z-value = -2.334, p-value = 0.0196

Yacé has a significant negative effect, but the magnitude of the effect is smaller than Bocou 1, Bonoua 34, and TMS30572. variety (Yavo vs Reference variety): Estimate = 0.82253, Std. Error = 0.75470, z-value = 1.090, p-value = 0.2758

The variety Yavo does not show a significant difference from the reference variety, indicating no clear evidence of its influence on cdm.

Key takeaways from the model: Whitefly count categories (wfCount_category):

Neither the medium nor high whitefly count categories show significant effects on the response variable cdm, suggesting that the whitefly count categories do not strongly influence the likelihood of cdm. Variety effects:

Several varieties (Bocou 1, Bonoua 34, TMS30572, Yacé) show a significant negative effect on the likelihood of cdm. This means that plants from these varieties are less likely to exhibit the outcome (cdm) compared to the reference variety.

The variety Yavo does not significantly differ from the reference variety in terms of its effect on cdm. Random effects:

There is some variability at both the plant and block levels, with the block effect having more variability (Std.Dev = 0.9113) compared to the plant effect (Std.Dev = 0.2475). This means that while individual plants contribute to variation, blocks introduce more variability in the outcome.

Conclusion:

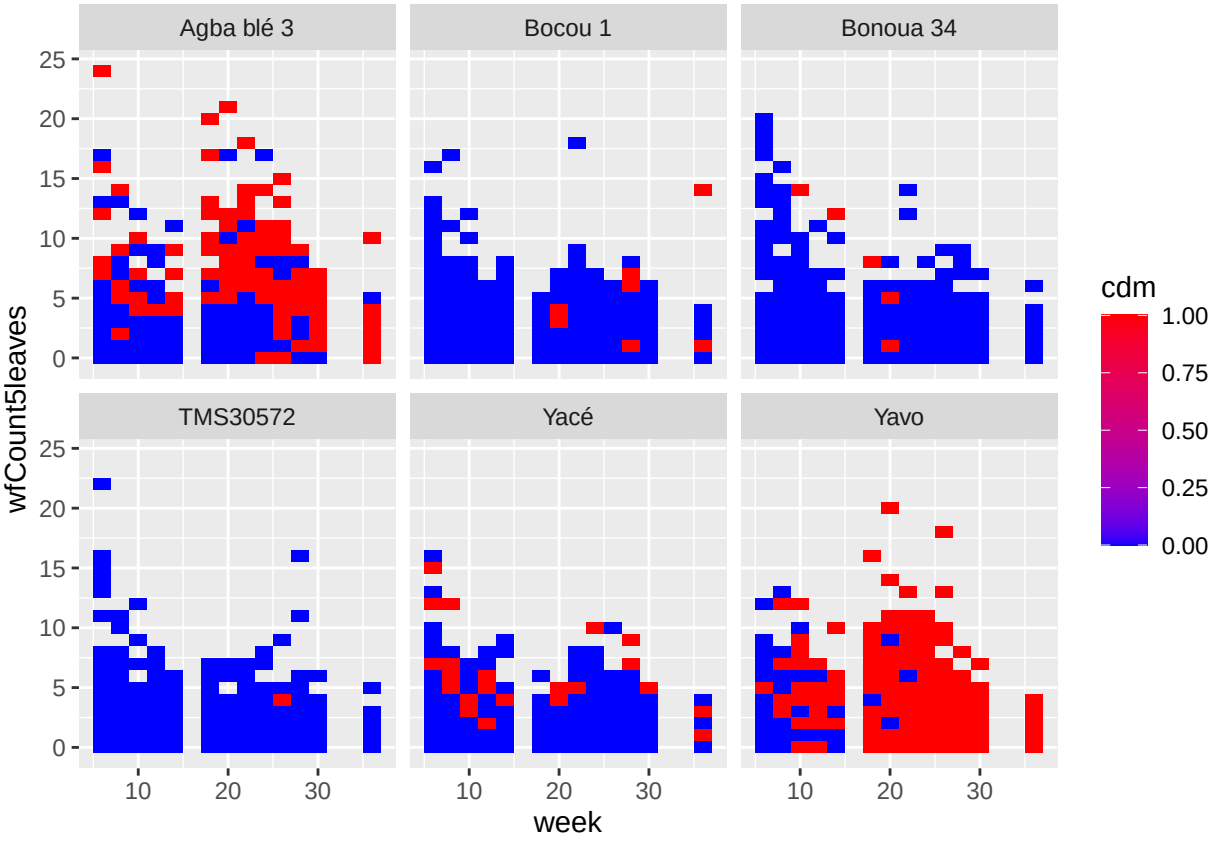
Variety plays a significant role in determining the likelihood of cdm. Specifically, certain varieties like Bocou 1, Bonoua 34, and TMS30572 have much lower odds of cdm compared to the reference variety.

Whitefly count does not seem to have a significant effect on the outcome, indicating that other factors (such as variety) may be more important in influencing cdm.

There is a non-negligible amount of unexplained variability at the plant and block levels, suggesting that further investigation into other potential sources of variation could be worthwhile.

Visualization

Heatmaps: Create heatmaps showing the relationship between whitefly counts and CMD incidence for each variety over time. This can visually reveal periods where increases in whitefly counts correspond to higher CMD incidence.



Objective 3: Impact of Resistant Varieties on Disease Dynamics

Disease Modeling: Develop a disease transmission model to simulate the spread of CMD in different cassava varieties.

Scenario Analysis: Evaluate the impact of resistant varieties on disease prevalence and spread under various scenarios.

This falls outside the primary scope of this analysis.

Objective 4: Disease Progression

Change in each variety

```
## # A tibble: 1,649 x 4
##    variety     week prev_severity severity
##    <chr>      <dbl>         <dbl>    <dbl>
##  1 Agba blé 3     6             1        2
##  2 Agba blé 3     6             2        1
##  3 Yacé           6             1        3
##  4 Yacé           6             3        1
##  5 TMS30572       6             1        2
##  6 TMS30572       6             2        1
##  7 Bonoua 34      6             1        2
##  8 Bonoua 34      6             2        1
##  9 Yavo           6             1        3
## 10 Yavo           6             3        2
## # i 1,639 more rows
```

```
##
## Attaching package: 'survival'
```

```
## The following object is masked from 'package:brms':
##
##     kidney
```

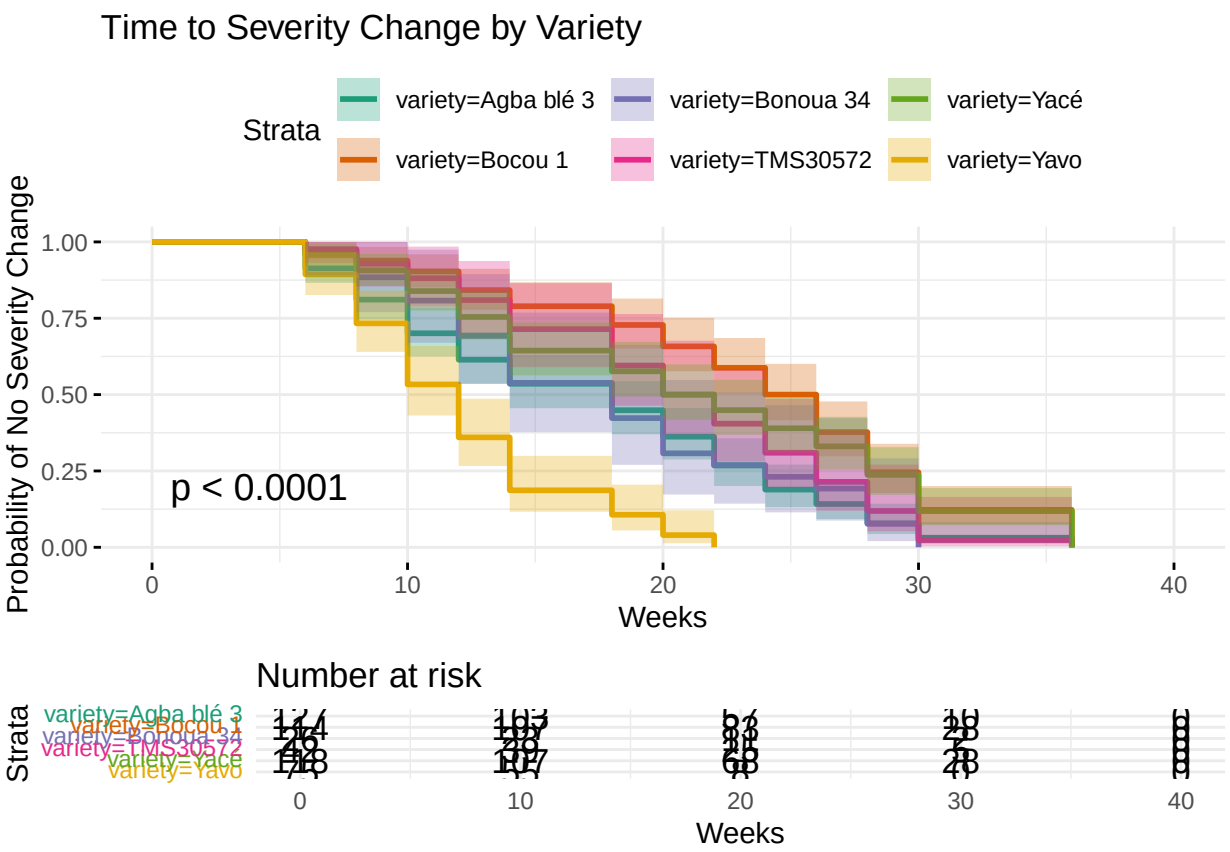
```
## Warning: package 'survminer' was built under R version 4.4.1
```

```
## Loading required package: ggpubr
```

```
##
## Attaching package: 'survminer'
```

```
## The following object is masked from 'package:survival':
##
##     myeloma
```

```
## Warning in Surv(time = score_changes$week, event = score_changes$severity):
## Invalid status value, converted to NA
```



Interpretation Agba blé 3: This variety tends to experience severity changes relatively early, with half of the plants showing a shift in score by 18 weeks. The interval for these changes ranges between 14 and 20 weeks, suggesting that Agba blé 3 is more susceptible to early progression of severity symptoms.

Bocou 1: With a median time of 26 weeks for severity changes, Bocou 1 shows the longest delay in symptom progression among the varieties. The 95% confidence interval, from 24 to 26 weeks, suggests this variety has a consistently slower progression of severity, making it a potentially more resilient choice where delayed symptom onset is desired.

Bonoua 34: Bonoua 34 exhibits severity changes with a median time of 18 weeks. The interval is broader, between 14 and 24 weeks, indicating some variability in how this variety responds, but it generally shows an intermediate rate of severity progression.

TMS30572: This variety has a median time of 21 weeks to severity change, with a range of 18 to 24 weeks. TMS30572 is slower to develop symptoms than Agba blé 3 but not as slow as Bocou 1, placing it in a moderate category for symptom progression.

Yacé: The Yacé variety has a median time of 24 weeks for severity changes, with an interval between 22 and 26 weeks. Similar to Bocou 1, this variety shows a delayed severity progression, though slightly more variable, which may be beneficial for longer resilience against symptom severity.

Yavo: Yavo experiences the fastest symptom progression, with a median severity change time of 12 weeks. The tight interval of 10 to 12 weeks indicates a very early and consistent onset of severity, making Yavo the most susceptible variety in this regard.