Reeb Biostats final

```
library(ggplot2)
library(tidyr)
library(knitr)
library(lme4)
library(lmerTest)
library(car)
library(dplyr)


options(contrasts = c("contr.sum", "contr.poly"))



### download indiv datasets of students
phenotesta = read.csv("~/Documents/Pitt Docs/Semester 3/iNaturalist/City_Family_Complete/Aster_PGH_CC_c

phenotestb = read.csv("~/Documents/Pitt Docs/Semester 3/iNaturalist/City_Family_Complete/Aster_PGH_edt_

phenotestc = read.csv("~/Documents/Pitt Docs/Semester 3/iNaturalist/City_Family_Complete/Aster_PHG_EO_c

### full join student datasets
phenotest = phenotesta %>%
  full_join(phenotestb, by = c( "observer", "id", "observed_on", "user_id", "url", "image_url", "latitu
  full_join(phenotestc, by =c(  "observer", "id", "observed_on", "user_id", "url", "image_url", "latitu
```

```
## fixing two messed up urls

phenotest$url[phenotest$id == 8020080] = "https://www.inaturalist.org/observations/8020080
"

phenotest$url[phenotest$id == 6759011] = "https://www.inaturalist.org/observations/6759011
"
```

```
## summarize by the number "distinct" identifiers for each observation (1 = all scorers agree, 2 = 2 sc
phenotest2 = phenotest %>%
  group_by(id, observed_on, url, scientific_name) %>%
  summarise(agree_part_or_whole = n_distinct(part_or_whole), agree_leaves = n_distinct(leaves), agree_f

## needed previously to find observation duplicates
##phenotest2$id[duplicated(phenotest2$id)]
```

```
## create a long-form dataset

phenotest2_long <- gather(phenotest2, phenophase, agreement, agree_part_or_whole:agree_ripe_fruit, fact

## PROBELM: how do I filter "obvious no" observations out of the late phenology stages? (ie only some p
```

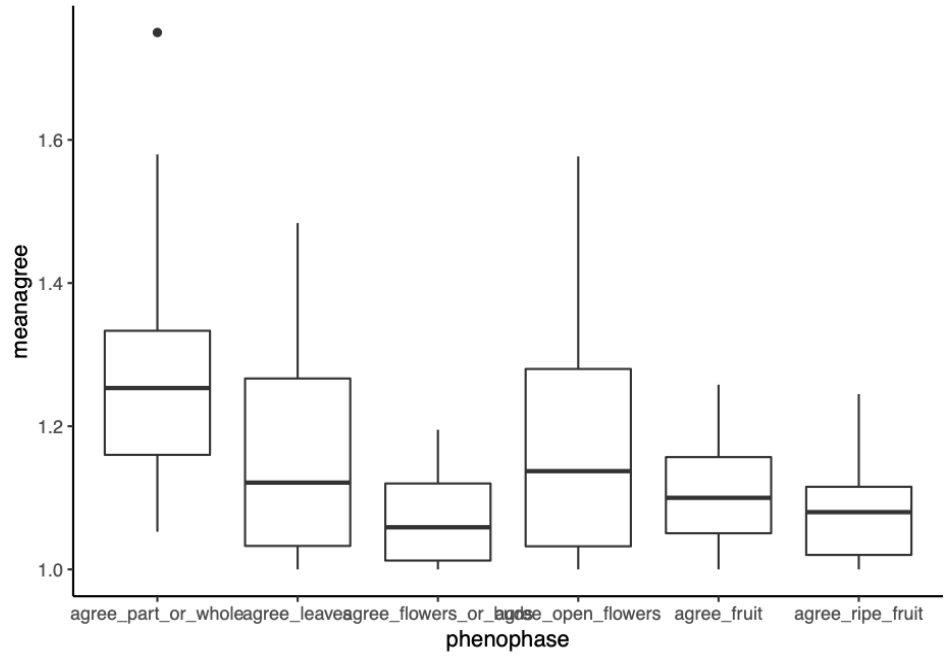
prelim data analysis

```
library(car)

phenotest2_long$agreement = as.numeric(phenotest2_long$agreement)


## type 3 anova test for unbalanced data
### NOT normally distributed - right skew
anovatest = aov(agreement ~ phenophase * scientific_name, data = phenotest2_long)
Anova(anovatest, type = 3)
# TukeyHSD(anovatest)

# both independently significant
kruskaltest_pheno = kruskal.test(agreement ~ phenophase, data = phenotest2_long)
kruskaltest_pheno


kruskaltest_species = kruskal.test(agreement ~ scientific_name, data = phenotest2_long)
kruskaltest_species
```

```
## Centering data by species?

phenotest2_long_sp = phenotest2_long %>%
  group_by(scientific_name, phenophase) %>%
  summarise(meanagree = mean(agreement))

aovtest = aov(meanagree ~ phenophase, data = phenotest2_long_sp)
anova(aovtest)
TukeyHSD(aovtest)
### after centering, only part or whole category is unique
```

boxplot for data centered by species

```
theme_set(theme_classic())

ggplot(data = phenotest2_long_sp, aes(x = phenophase, y = meanagree)) + geom_boxplot() + scale_fill_bre
```



making new binomial dataset

```
phenotest2_long$agreement = as.numeric(phenotest2_long$agreement)

attach(phenotest2_long)

phenotest2_long$disagree_binary[phenotest2_long$agreement >= 2] = 1
```

```
## Warning: Unknown or uninitialised column: 'disagree_binary'.
```

```
phenotest2_long$disagree_binary[phenotest2_long$agreement <= 1] = 0

detach(phenotest2_long)
```

———FINAL———

final logistic regression test:

```r
logtest = glm(disagree_binary ~ phenophase + scientific_name, data = phenotest2_long, family = binomial
summary(logtest)
```

```
##
## Call:
## glm(formula = disagree_binary ~ phenophase + scientific_name,
##     family = binomial(link = "logit"), data = phenotest2_long)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0435  -0.5942  -0.4442  -0.3116   2.7327
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -1.95311    0.05195 -37.598  < 2e-16 ***
## phenophase1        1.01534    0.07915  12.828  < 2e-16 ***
## phenophase2        0.19619    0.09296   2.111 0.034805 *
## phenophase3       -0.78457    0.12716  -6.170 6.83e-10 ***
## phenophase4        0.24436    0.09183   2.661 0.007789 **
## phenophase5       -0.13303    0.10191  -1.305 0.191777
## scientific_name1   0.15634    0.13891   1.125 0.260383
## scientific_name2  -0.23866    0.14552  -1.640 0.100999
## scientific_name3   0.18194    0.21579   0.843 0.399150
## scientific_name4   0.26180    0.17052   1.535 0.124697
## scientific_name5   0.61433    0.13023   4.717 2.39e-06 ***
## scientific_name6   0.09057    0.13639   0.664 0.506646
## scientific_name7   0.23127    0.15568   1.486 0.137410
## scientific_name8   0.57432    0.22012   2.609 0.009077 **
## scientific_name9   0.07208    0.22723   0.317 0.751105
## scientific_name10 -0.48175    0.30592  -1.575 0.115304
## scientific_name11  0.11216    0.23861   0.470 0.638326
## scientific_name12  0.34749    0.22794   1.524 0.127391
## scientific_name13 -0.27942    0.16759  -1.667 0.095462 .
## scientific_name14 -0.51667    0.25464  -2.029 0.042458 *
## scientific_name15 -0.97205    0.17257  -5.633 1.77e-08 ***
## scientific_name16 -0.50924    0.14832  -3.433 0.000596 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4022.5  on 5081  degrees of freedom
## Residual deviance: 3720.6  on 5060  degrees of freedom
## AIC: 3764.6
##
## Number of Fisher Scoring iterations: 5
```

```r
anova(logtest, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
```

```
##
## Response: disagree_binary
##
## Terms added sequentially (first to last)
##
##
##                 Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                             5081     4022.5
## phenophase       5   196.43      5076     3826.0 < 2.2e-16 ***
## scientific_name 16   105.45      5060     3720.6 3.265e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
### anova comparing this model to a null model finds that phenophase and species were significantly dif
```

final frequency table of binary agreements

```
table_pheno = phenotest2_long %>%
  dplyr::group_by(phenophase, disagree_binary) %>%
  dplyr::summarise(n = n()) %>%
  dplyr::mutate(freq = n / sum(n))
table_pheno
```

```
## # A tibble: 12 x 4
## # Groups:   phenophase [6]
##    phenophase            disagree_binary     n   freq
##    <fct>                           <dbl> <int>  <dbl>
##  1 agree_part_or_whole                 0   617 0.728
##  2 agree_part_or_whole                 1   230 0.272
##  3 agree_leaves                        0   725 0.856
##  4 agree_leaves                        1   122 0.144
##  5 agree_flowers_or_buds               0   796 0.940
##  6 agree_flowers_or_buds               1    51 0.0602
##  7 agree_open_flowers                  0   720 0.850
##  8 agree_open_flowers                  1   127 0.150
##  9 agree_fruit                         0   755 0.891
## 10 agree_fruit                         1    92 0.109
## 11 agree_ripe_fruit                    0   783 0.924
## 12 agree_ripe_fruit                    1    64 0.0756
```

```
table_pheno2 = phenotest2_long %>%
  dplyr::group_by(scientific_name, disagree_binary) %>%
  dplyr::summarise(n = n()) %>%
  dplyr::mutate(freq = n / sum(n))
table_pheno2
```

```
## # A tibble: 34 x 4
## # Groups:   scientific_name [17]
##    scientific_name         disagree_binary     n  freq
##    <fct>                             <dbl> <int> <dbl>
```

```
##  1 Achillea millefolium                0   349 0.843
##  2 Achillea millefolium                1    65 0.157
##  3 Ageratina altissima                 0   431 0.887
##  4 Ageratina altissima                 1    55 0.113
##  5 Ambrosia artemisiifolia             0   131 0.840
##  6 Ambrosia artemisiifolia             1    25 0.160
##  7 Artemisia vulgaris                  0   204 0.829
##  8 Artemisia vulgaris                  1    42 0.171
##  9 Cichorium intybus                   0   289 0.777
## 10 Cichorium intybus                   1    83 0.223
## # ... with 24 more rows
```

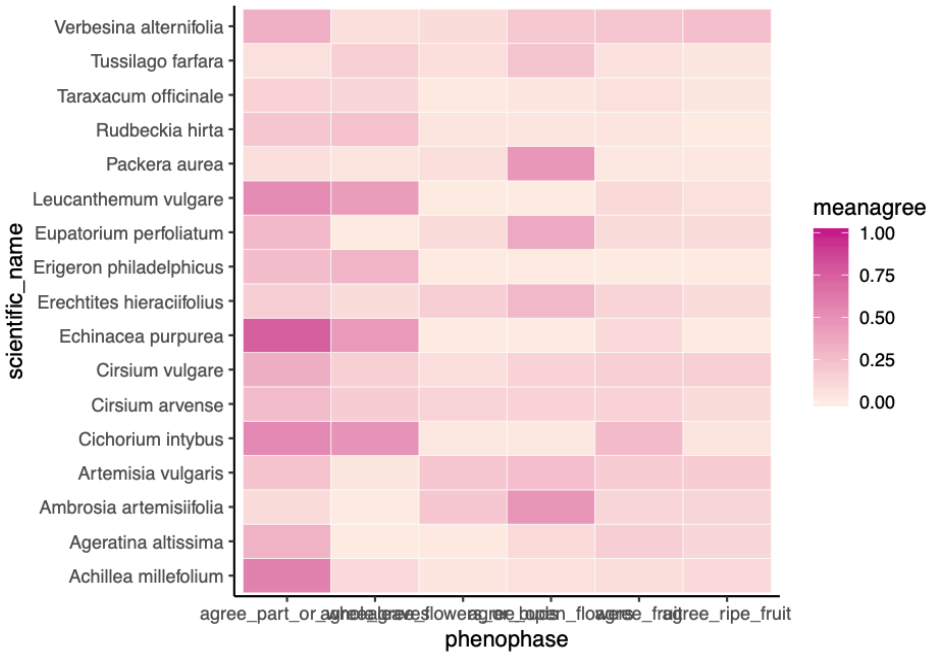
```
table_pheno3 = phenotest2_long %>%
  dplyr::group_by(disagree_binary) %>%
  dplyr::summarise(n = n()) %>%
  dplyr::mutate(freq = n / sum(n))
table_pheno3
```

```
## # A tibble: 2 x 3
##   disagree_binary     n  freq
##             <dbl> <int> <dbl>
## 1               0  4396 0.865
## 2               1   686 0.135
```

Final Figure

```
meangroup = phenotest2_long %>%
  dplyr::group_by(scientific_name, phenophase) %>%
  dplyr::summarise(meanagree = mean(disagree_binary))
```

```
theme_set(theme_classic())

ggplot(meangroup, aes(phenophase, scientific_name, fill = meanagree)) + geom_tile(color = "white") +
  scale_fill_gradient(limits = c(0,1), low="#feebe2", high="#c51b8a")
```


Verbesina alternifolia
Tussilago farfara
Taraxacum officinale
Rudbeckia hirta
Packera aurea
Leucanthemum vulgare
Eupatorium perfoliatum
Erigeron philadelphicus
Erechtites hieraciifolius
Echinacea purpurea
Cirsium vulgare
Cirsium arvense
Cichorium intybus
Artemisia vulgaris
Ambrosia artemisiifolia
Ageratina altissima
Achillea millefolium
scientific_name
agree_part_or_all_leaves
agree_ripe_fruit
phenophase
meanagree
1.00
0.75
0.50
0.25
0.00