2-combined-analysis

IndoorOutdoor Team

6/08/2017

Report

Background/Justification

- Farmers are often interested in knowing the growth rate for sorghum, so that they can better predict harvest time.
- It is much easier to gather data on sorghum grown indoors compared to outdoors.
- The TERRA-REF database includes information on sorghum grown in greenhouses as well as sorghum grown outdoors.
- Our goal is to use this data to find a preliminary relationship between the growth rate for sorghum grown indoors versus sorghum grown outdoors.

Data Description

Step 1: Pull data from the TERRA-REF database * traits + table with all metrics gathered * variables + Contains list of all metrics/data gathered in the TERRA-REF project * cultivars + cultivar * entities + Contains list of machines used to collect data + Unused * sites + Contains list of location of plants * treatments + Contains list of treatments for plants + Unused * experiments + Contains list of all experiments performed, including the planting date for different seasons

Step 2: Join TERRA-REF tables together to have a table with all variables relevant to our project * Joined_table + Joined together traits, variables, cultivars, entities, sites, treatments tables by the appropriate ids

Step 3: Filter table for records we need and add new variables as needed * filtered_table + Filtered the joined_table for the metrics that best describe growth, including height, leaf_width, leaf_length * height_table + Further analysis showed that the only metric with data from both indoor and outdoor sorghum is height + Filtered the filtered_table for three metrics: height, plant_height, canopy_height + Filtered out any records without date + Added columns - indoor_outdoor (whether the metric is taken indoor or outdoor, based on site variable) - height_cm (height, plant_height, canopy_height had different units of measurement; this column changes all to cm) - Date (takes the day, month, year information from date and constructs a date object) - season (uses Date to categorize plant into seasons) - age (found start date for each season in experiments table and from Dr. LeBauer, calculated age of plant in days) - site_season (concatenates indoor_outdoor and season values for plotting purposes) * indoor_height + Table consisting of only the indoor height data * outdoor_height + Table consisting of only the outdoor height data

Approach & Methods

- Our initial thinking
- In the ideal case, our goal is with all the other conditions fixed, to look at the difference between heights of indoor and outdoor plants. Specifically we can do it with all the available predictors like cultivars and treatments, and compare different variables such as canopy heights, canopy cover, leaf lengths and leaf widths.
- Our first step included plotting the data (height again age) to get a sense for the data

- Unfortunately, the data was far from ideal. The only variable shared by indoor and outdoor plants is height (plant heights for indoor plants and canopy heights for outdoor ones). Also, the indoor plants and outdoor plants were grown in different places and different times, and neither the cultivars nor the treatment types of the indoor plants coincided with those of the outdoor plants. As a consequence, it's not feasible to control variables.
- Anyway, what we can do is try to build a model to show the relation between indoor and outdoor heights. We first looked at the indoor and outdoor heights data separately. We made plots of them and set up our first linear models. For the indoor model we found cultivar is a significant factor, but treatment type is not. For the outdoor model, we found season to be a significant factor.
- After that we put the indoor and outdoor data together. This time we could only include a factor specifying indoor or outdoor, and the model showed that indoor plants grow much faster. However, the linear models were not quite satisfying, since they predicted negative values for heights at 0 age. Also, the variances seemed to be too large. Probably we could fix this by tracing the growing history of every single plant.

Analysis & Results

- A table that consists of only the Indoor data is created.
- The significant variables from the indoor table have been considered for modeling. The interaction terms for each predictor variable have been considered.
- The QQ plot for each model was observed.
- The models were made better by removing the influential and the leverage points. The predictor variable was also transformed to ensure that the model follows the Normality Assumption.
- Box-Cox transformation was implemented to obtain a better model.
- The final indoor model that we chose is:
- A similar approach was followed for Outdoor data. A separate table that consists of just the Outdoor data was created.
- The rows with the missing date values were not considered for modeling.
- The final outdoor model that we chose is:

Discussion

- Discussion of findings
- We suspect that these findings will not be surprising to anyone. Plants grown in a manufactured environment grow more quickly than those subject to unpredictable weather patterns. Since we only had three crops of plants, our findings cannot be very conclusive. We are able to propose, however, some suggestions for further research into this topic.
- Ideas for future exploration:
- If we were to continue this project, we would want more data to work with, particularly plant growth from more seasons and more outdoor locations. This way, we could draw more general conclusions about outdoor sorghum growth. Further, we would like to have more measures of growth to analyze, such as plant diameter, leaf width, and crop yield. These would give a more full picture of the overall growth than merely plant height.
- We also wanted to explore growth rates of particular cultivars of sorghum bicolor, but ran out of time to complete this analysis. Further, there were very few cultivars which were grown both indoor and outdoor. For a full analysis, it would be helpful to have data about the growth of more cultivars, so we could determine if different cultivars fare better in different conditions.
- Finally, we would want to incorporate weather data in our analysis. We suspect that plants grown outdoors in conditions more similar to the greenhouse environment might yield higher growth.

Data Cleaning & Filtering

Loading BetyDB into environment and joining to make main table

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
bety_src <- src_postgres(dbname = "bety",
                password = 'bety',
                host = 'terra-bety.default',
                user = 'bety',
                port = 5432)

# to see all available columns in traits table
original_traits <- tbl(bety_src, 'traits') %>%
  collect(n=1)
```

```
## Warning: Only first 1 results retrieved. Use n = Inf to retrieve all.
```

```r
# local version of variables for reference
variables_local <- tbl(bety_src, 'variables', n = Inf) %>%
  mutate(variable_id = id, variable_name = name) %>%
  dplyr::select(variable_id, variable_name, description, units) %>%
  collect()

traits <- tbl(bety_src, 'traits', n = Inf) %>%
  mutate(trait_id = id) %>%
  dplyr::select(trait_id, site_id, specie_id, cultivar_id, date, mean, variable_id, method_id, treatmen

variables <- tbl(bety_src, 'variables', n = Inf) %>%
  mutate(variable_id = id, variable_name = name) %>%
  dplyr::select(variable_id, variable_name)

cultivars <- tbl(bety_src, 'cultivars', n = Inf) %>%
  mutate(cultivar_id = id, cultivar = name) %>%
  dplyr::select(cultivar_id, cultivar)

entities <- tbl(bety_src, 'entities', n = Inf) %>%
  mutate(entity_name = name, entity_id = id) %>%
  dplyr::select(entity_name, entity_id)

sites <- tbl(bety_src, 'sites', n = Inf) %>%
  mutate(site_id = id) %>%
  dplyr::select(site_id, city, state, country, notes, sitename)
```

```
## Warning in postgresqlExecStatement(conn, statement, ...): RS-DBI driver
## warning: (unrecognized PostgreSQL field type geometry (id:20504) in column
## 17)
```

```
treatments <- tbl(bety_src, 'treatments', n = Inf) %>%
  mutate(treatment_id = id, treatment_definition = definition, treatment_name = name) %>%
  dplyr::select(treatment_id, treatment_name, treatment_definition)

# looking for when each season began
experiments <- tbl(bety_src, 'experiments', n = Inf) %>%
  #mutate(treatment_id = id, treatment_definition = definition, treatment_name = name) %>%
  #dplyr::select(treatment_id, treatment_name, treatment_definition)
  collect(n = Inf)

# join relevant tables together
joined_table <- traits %>%
  left_join(variables, by = 'variable_id') %>%
  left_join(cultivars, by = 'cultivar_id') %>%
  left_join(entities, by = 'entity_id') %>%
  left_join(sites, by = 'site_id') %>%
  left_join(treatments, by = 'treatment_id') %>%
  dplyr::select(trait_id, date, mean, variable_name, sitename, treatment_name, cultivar)
```

Filtering Table and Definining New Variables

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following object is masked from 'package:base':
##
##     date
```

```
# filter table for measurements that we care about
filtered_table <- filter(joined_table, variable_name %in% c("height", "canopy_cover", "canopy_height",

height_table <- filter(filtered_table, variable_name %in% c("canopy_height", "plant_height", "height"))
  collect(n = Inf)

# filter table for rows without date
height_table <- filter(height_table, !is.na(height_table$date))

# define variable indoor_outdoor based on sitename
height_table$indoor_outdoor <- ifelse(height_table$sitename == 'Danforth Plant Science Center Bellweathe

# define variable height(cm)
# canopy_height is in , plant_height is in
# range(filter(height_table, height_table$variable_name == "canopy_height")$mean) gives 5 and 270
# range(filter(height_table, height_table$variable_name == "plant_height")$mean) gives 17 and 1671, var
height_table$height_cm <- if_else(height_table$variable_name == "canopy_height", height_table$mean, if_e

# define variable Date that is a date type
height_table$Date <- as.Date(substr(height_table$date, 0, 10), format = "%Y-%m-%d")
```

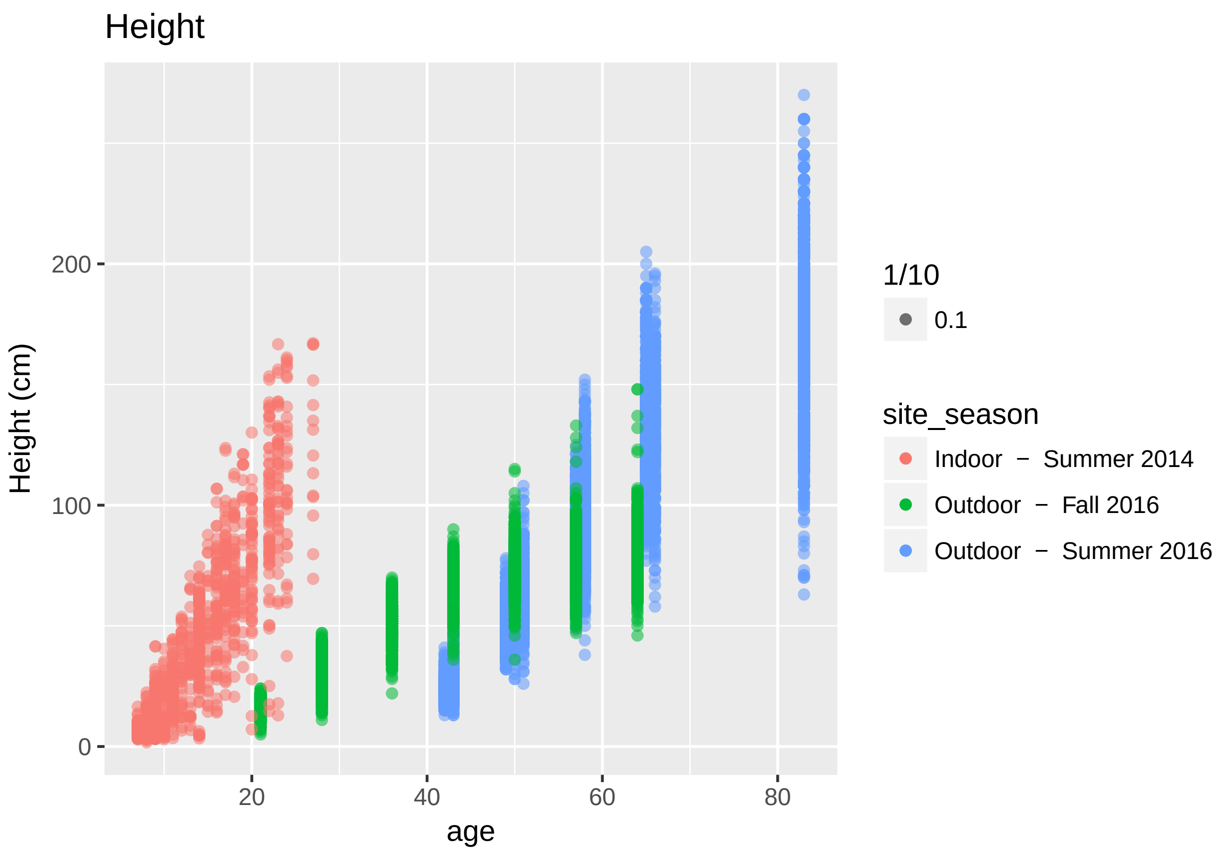
```
# define new variable season based on which season the data was taken
summer2014 = interval(as.Date("2014-05-01", format = "%Y-%m-%d"), as.Date("2014-07-31", format = "%Y-%m-
summer2016 = interval(as.Date("2016-05-01", format = "%Y-%m-%d"), as.Date("2016-07-31", format = "%Y-%m-
fall2016 = interval(as.Date("2016-08-15", format = "%Y-%m-%d"), as.Date("2016-12-31", format = "%Y-%m-%

# checked whether all data in height_table fall into one of three seasons defined
height_table$season <- if_else(height_table$Date %within% summer2014, "Summer 2014", if_else(height_tabl

# define new variable age based on when the season began
# 2016 season start dates came from experiments table, 2014 season start date came from Dr. LeBauer
summer2014start = as.Date("2014-05-27", format = "%Y-%m-%d")
summer2016start = as.Date("2016-04-19", format = "%Y-%m-%d")
fall2016start = as.Date("2016-08-03", format = "%Y-%m-%d")

height_table$age <- if_else(height_table$season == "Summer 2014", as.numeric(height_table$Date - summer2

height_table$site_season <- paste(height_table$indoor_outdoor, " - ", height_table$season)

canopy_height_plant_height <- filter(height_table, variable_name %in% c("canopy_height", "plant_height")
  collect(n = Inf)

# look at which sites have which variables and cultivars
site_variable <- height_table %>%
  group_by(indoor_outdoor, variable_name, cultivar) %>%
  summarize(n = n()) %>%
  collect(n = Inf)

shared_cultivar <- filter(site_variable, cultivar %in% c('BTx642', 'PI_564163', 'Tx430', 'TX7000')) %>%
  collect(n = Inf)
```

Plotting

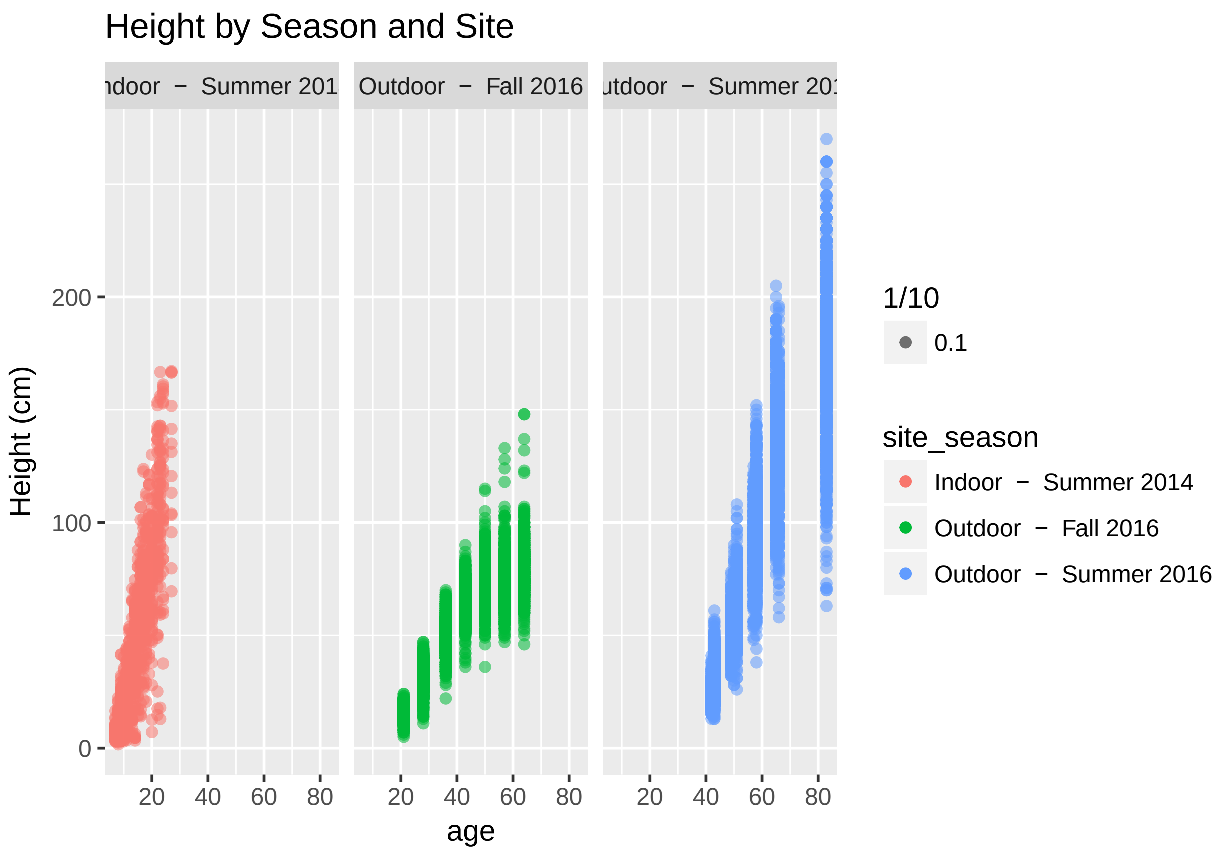
```
library(ggplot2)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine
```

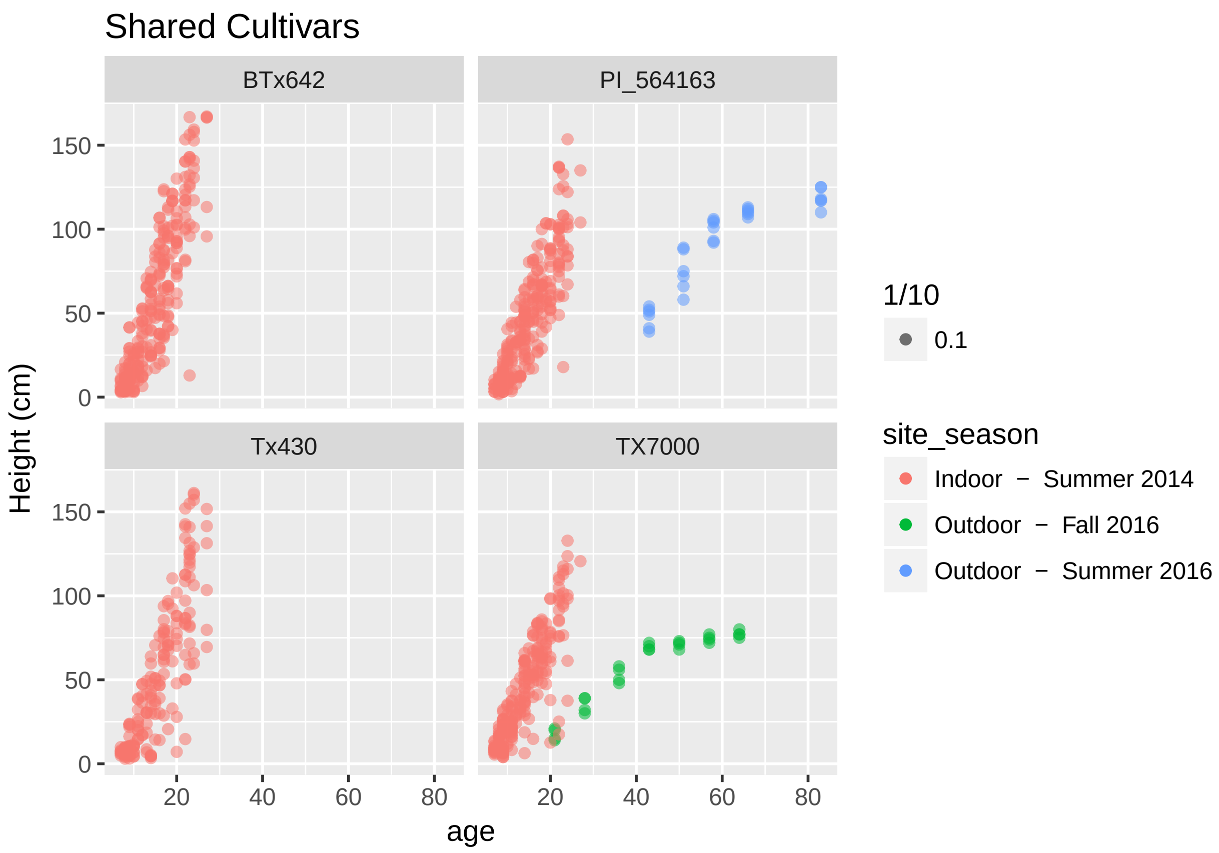
```
# plotting height against age
ggplot(canopy_height_plant_height, aes(x = age, y = height_cm)) +
  geom_point(aes(color = site_season, alpha = 1/10)) +
  ylab("Height (cm)") +
  ggtitle("Height")
```



```
ggsave("height_vs_age.png", width = 13, height = 7, path = "~/IndoorOutdoor/data")

ggplot(canopy_height_plant_height, aes(x = age, y = height_cm)) +
  geom_point(aes(color = site_season, alpha = 1/10)) +
  facet_wrap(~site_season) +
  ylab("Height (cm)") +
  ggtitle("Height by Season and Site")
```



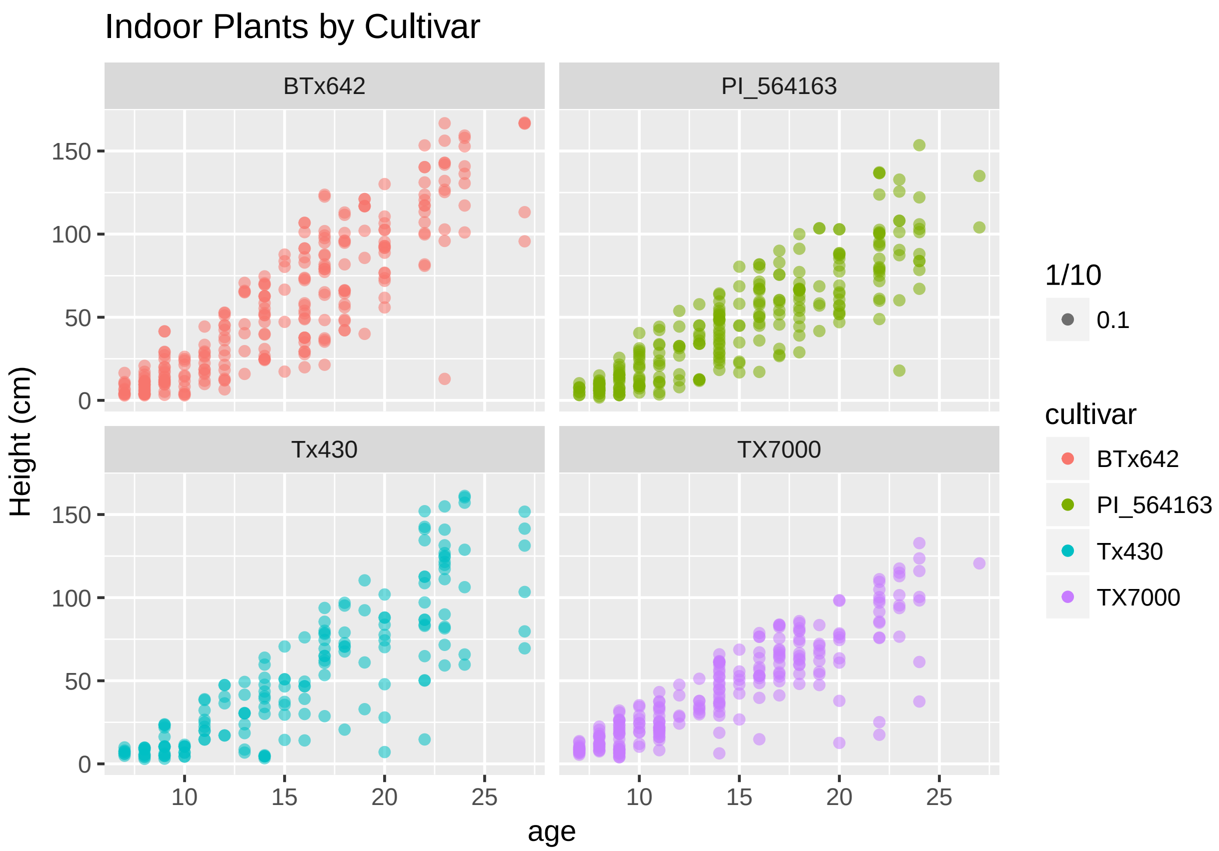
```
ggsave("height_by_season_and_site.png", width = 15, height = 3.8, path = "~/IndoorOutdoor/data")

# shared cultivars between indoor and outdoor

ggplot(filter(canopy_height_plant_height, cultivar %in% c('BTx642', 'PI_564163', 'Tx430', 'TX7000')), a
  geom_point(aes(color = site_season, alpha = 1/10)) +
  facet_wrap(~cultivar) +
  ylab("Height (cm)") +
  ggtitle("Shared Cultivars")
```



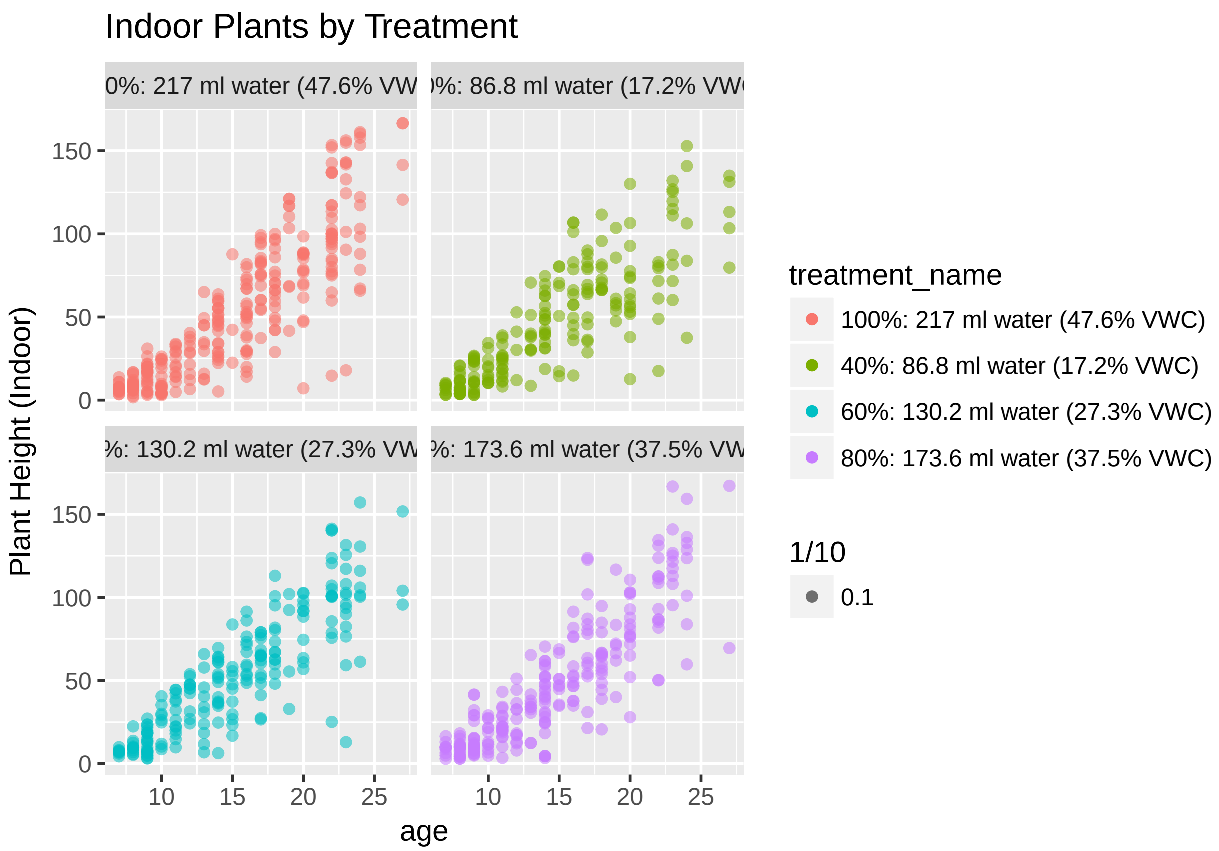
```
ggsave("shared_cultivars.png", path = "~/IndoorOutdoor/data")
```

```
## Saving 6.5 x 4.5 in image
```

```
# indoor plants
indoor_height <- filter(canopy_height_plant_height, indoor_outdoor == 'Indoor')

ggplot(indoor_height, aes(x = age, y = height_cm)) +
  geom_point(aes(color = cultivar, alpha = 1/10)) +
  facet_wrap(~cultivar) +
  ylab("Height (cm)") +
  ggtitle("Indoor Plants by Cultivar")
```



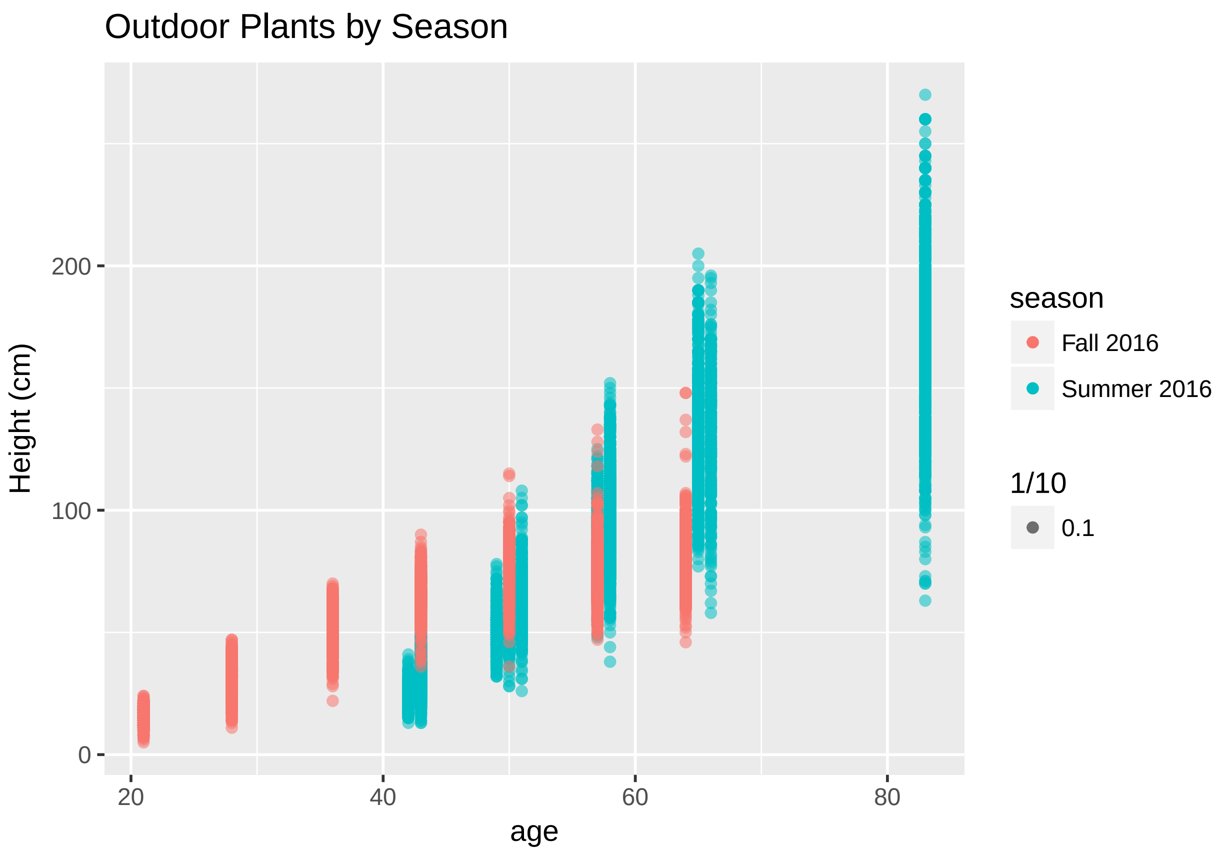
```
ggsave("indoor_plants_by_cultivar.png", height = 5, width = 7, path = "~/IndoorOutdoor/data")

ggplot(indoor_height, aes(x = age, y = height_cm)) +
  geom_point(aes(color = treatment_name, alpha = 1/10)) +
  facet_wrap(~treatment_name) +
  ylab("Plant Height (Indoor)") +
  ggtitle("Indoor Plants by Treatment")
```



```
ggsave("indoor_plants_by_treatment.png", path = "~/IndoorOutdoor/data")
```

```
## Saving 6.5 x 4.5 in image
```

```
# outdoor plants
outdoor_height <- filter(canopy_height_plant_height, indoor_outdoor == 'Outdoor')

ggplot(outdoor_height, aes(x = age, y = height_cm)) +
  geom_point(aes(color = season, alpha = 1/10)) +
  ylab("Height (cm)") +
  ggtitle("Outdoor Plants by Season")
```



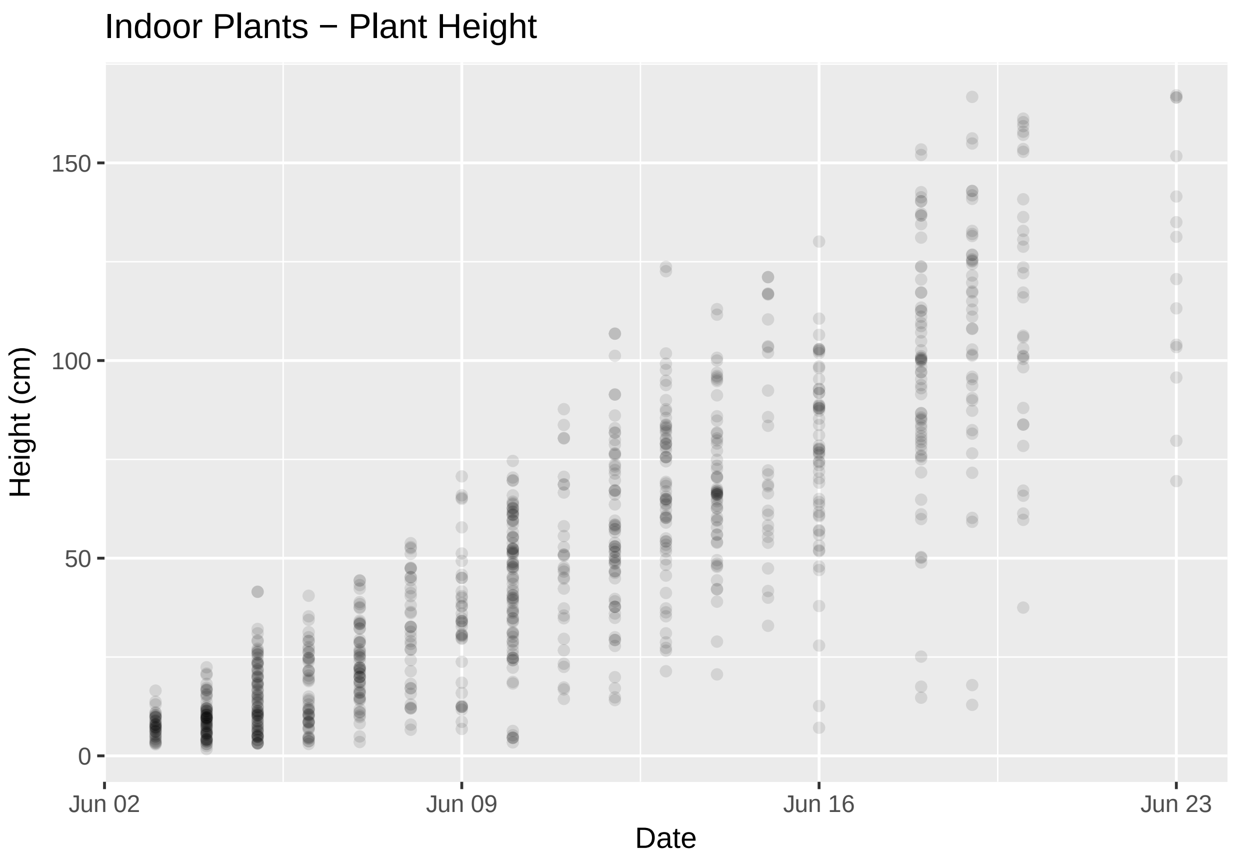
```
ggsave("outdoor_plants_by_season.png", path = "~/IndoorOutdoor/data")
```

```
## Saving 6.5 x 4.5 in image
```

Modeling - Emily

Plots

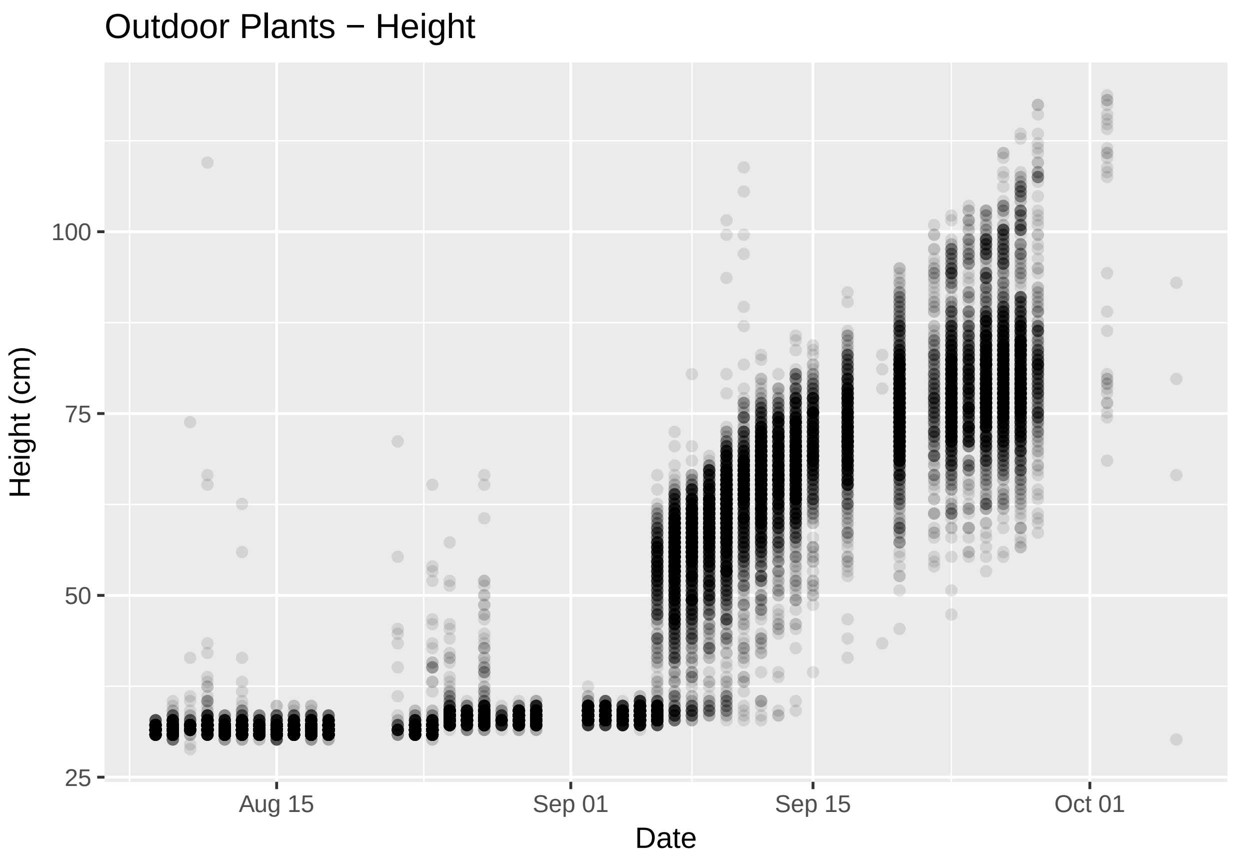
```
ggplot(indoor_height, aes(x = Date, y = height_cm)) +
  geom_point(alpha = 1/10) +
  ylab("Height (cm)") +
  ggtitle("Indoor Plants - Plant Height")
```



```r
ggsave("indoor_plants_plant_height.png", path = "~/IndoorOutdoor/data")
```

```
## Saving 6.5 x 4.5 in image
```

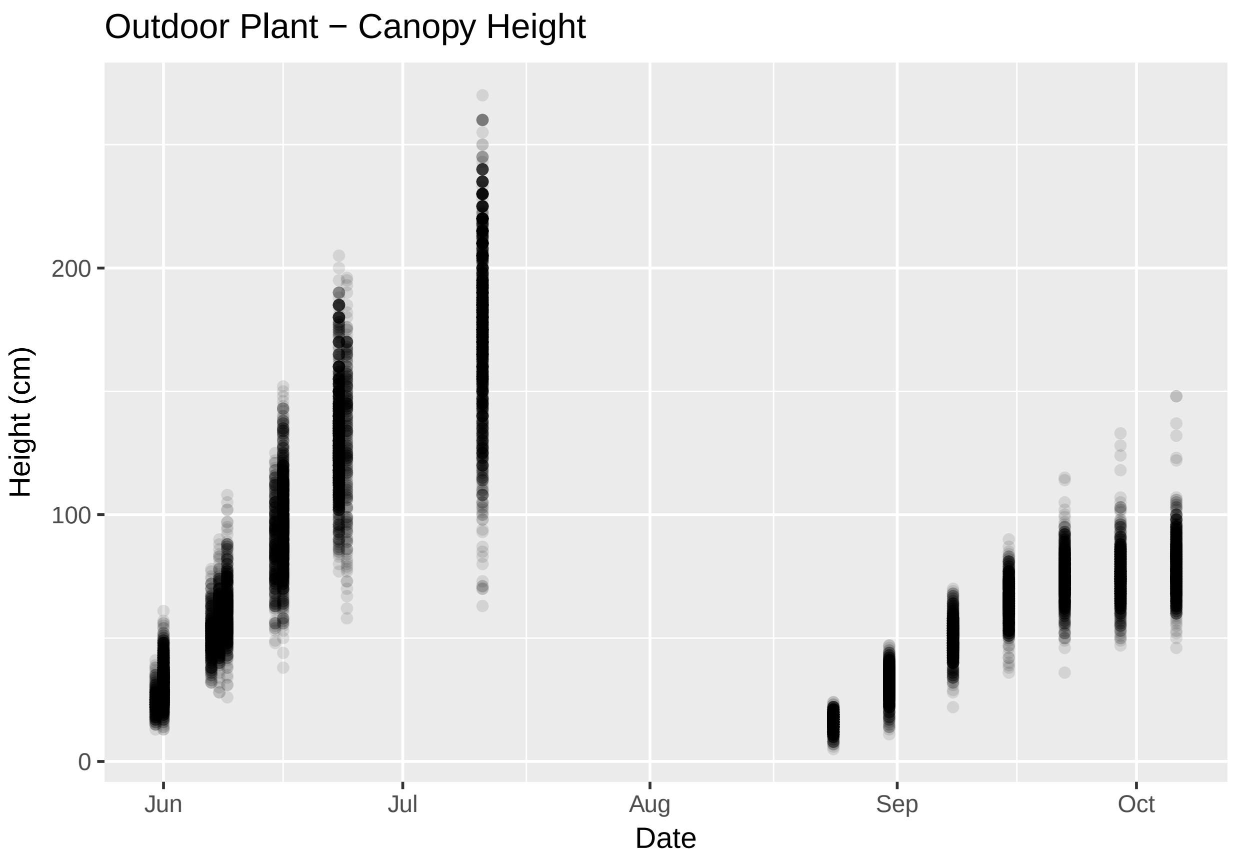
```r
ggplot(filter(height_table, variable_name == "height"), aes(x = Date, y = height_cm)) +
  geom_point(alpha = 1/10) +
  ylab("Height (cm)") +
  ggtitle("Outdoor Plants - Height")
```



```
ggsave("outdoor_plants_height.png", path = "~/IndoorOutdoor/data")
```

```
## Saving 6.5 x 4.5 in image
```

```
ggplot(filter(height_table, variable_name == "canopy_height"), aes(x = Date, y = height_cm)) +
  geom_point(alpha = 1/10) +
  ylab("Height (cm)") +
  ggtitle("Outdoor Plant - Canopy Height")
```



```
ggsave("outdoor_plants_canopy_height.png", path = "~/IndoorOutdoor/data")
```

```
## Saving 6.5 x 4.5 in image
```

Linear Regression!

```
indoor_growthmodel <- lm(height_cm ~ age, data = indoor_height)
summary(indoor_growthmodel)
```

```
##
## Call:
## lm(formula = height_cm ~ age, data = indoor_height)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -88.909  -9.799   0.211  10.092  64.891
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -43.1856     2.0096  -21.49   <2e-16 ***
## age           6.3041     0.1272   49.58   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.07 on 913 degrees of freedom
## Multiple R-squared:  0.7291, Adjusted R-squared:  0.7288
## F-statistic:  2458 on 1 and 913 DF,  p-value: < 2.2e-16
```

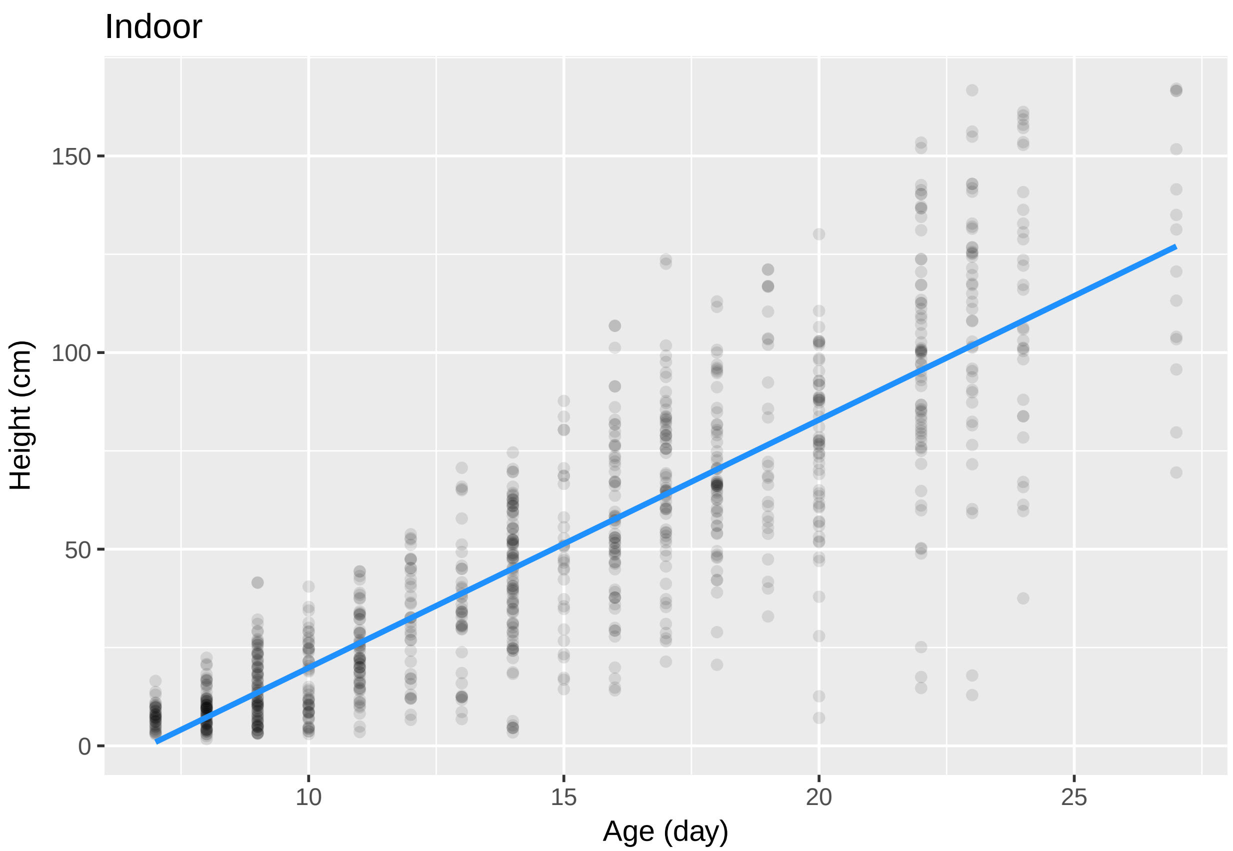
```
outdoorsummer_growthmodel <- lm(height_cm ~ age, data = outdoor_height[outdoor_height$season == "Summer
summary(outdoorsummer_growthmodel)
```

```
##
## Call:
## lm(formula = height_cm ~ age, data = outdoor_height[outdoor_height$season ==
##     "Summer 2016", ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -122.525  -10.400   -1.697   10.081   87.031
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -125.98294    1.12586  -111.9   <2e-16 ***
## age            3.75310    0.01849   203.0   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.12 on 6991 degrees of freedom
## Multiple R-squared:  0.8549, Adjusted R-squared:  0.8549
## F-statistic: 4.12e+04 on 1 and 6991 DF,  p-value: < 2.2e-16
```

```
outdoorfall_growthmodel <- lm(height_cm ~ age, data = outdoor_height[outdoor_height$season == "Fall 201
summary(outdoorfall_growthmodel)
```

```
##
## Call:
## lm(formula = height_cm ~ age, data = outdoor_height[outdoor_height$season ==
##     "Fall 2016", ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -42.314  -7.093  -0.771   7.229  59.686
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -8.07869    0.47028  -17.18   <2e-16 ***
## age          1.50614    0.01044  144.30   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.49 on 4896 degrees of freedom
## Multiple R-squared:  0.8096, Adjusted R-squared:  0.8096
## F-statistic: 2.082e+04 on 1 and 4896 DF,  p-value: < 2.2e-16
```

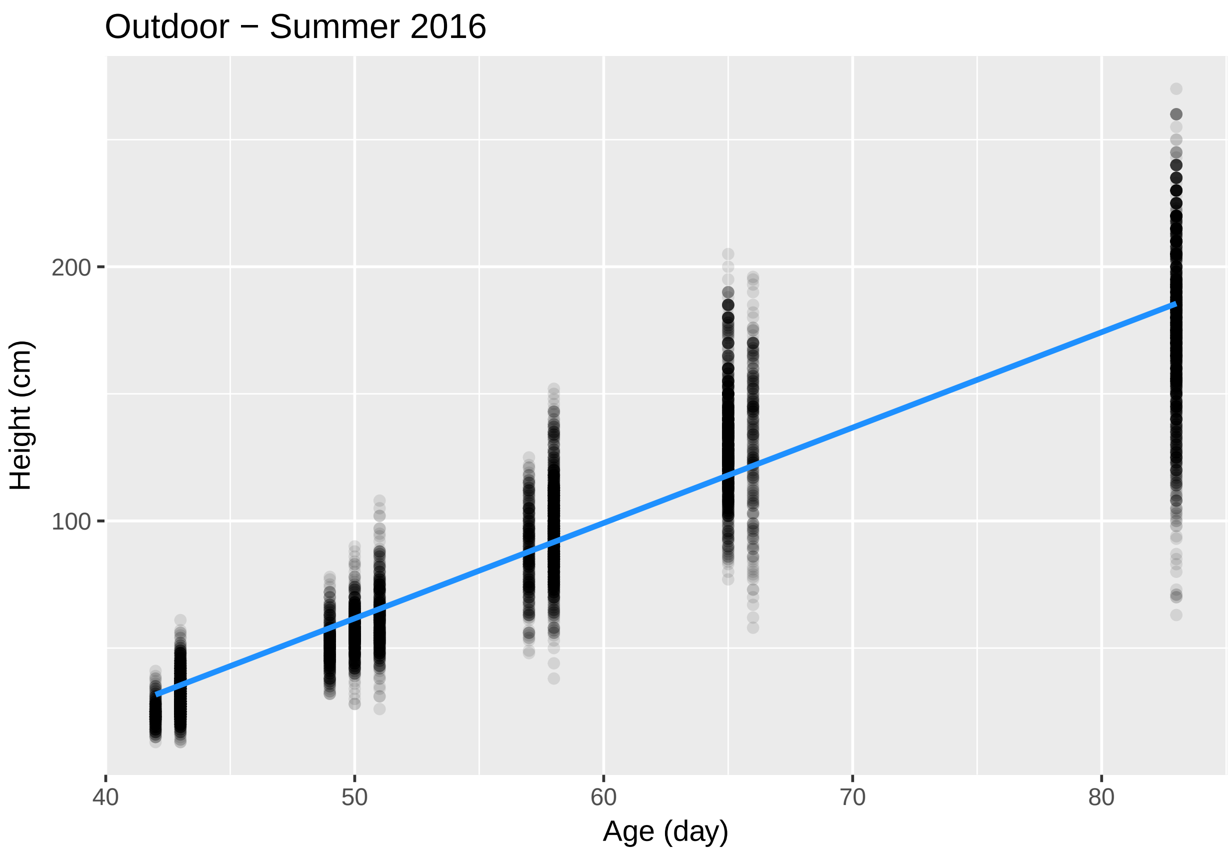
```
ggplot(indoor_height, aes(x = age, y = height_cm)) +
  geom_point(alpha = 1/10) +
  geom_smooth(method = "lm", col = "dodgerblue", se = FALSE) +
  ylab("Height (cm)") +
  xlab("Age (day)") +
  ggtitle("Indoor")
```



```
ggsave("indoor_model.png", path = "~/IndoorOutdoor/data")
```

```
## Saving 6.5 x 4.5 in image
```

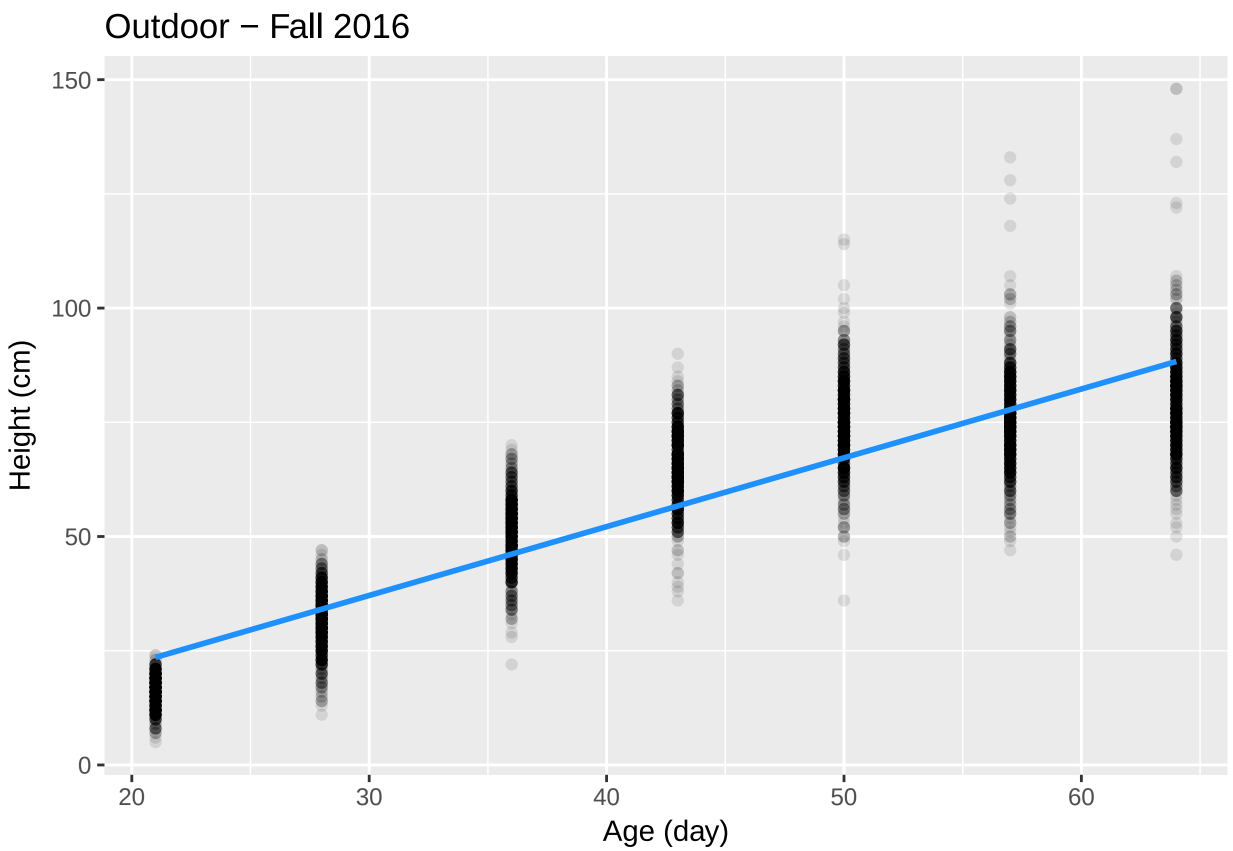
```
ggplot(filter(outdoor_height, season == "Summer 2016"), aes(x = age, y = height_cm)) +
  geom_point(alpha = 1/10) +
  geom_smooth(method = "lm", col = "dodgerblue", se = FALSE) +
  ylab("Height (cm)") +
  xlab("Age (day)") +
  ggtitle("Outdoor - Summer 2016")
```



```
ggsave("outdoor_summer_model.png", path = "~/IndoorOutdoor/data")
```

```
## Saving 6.5 x 4.5 in image
```

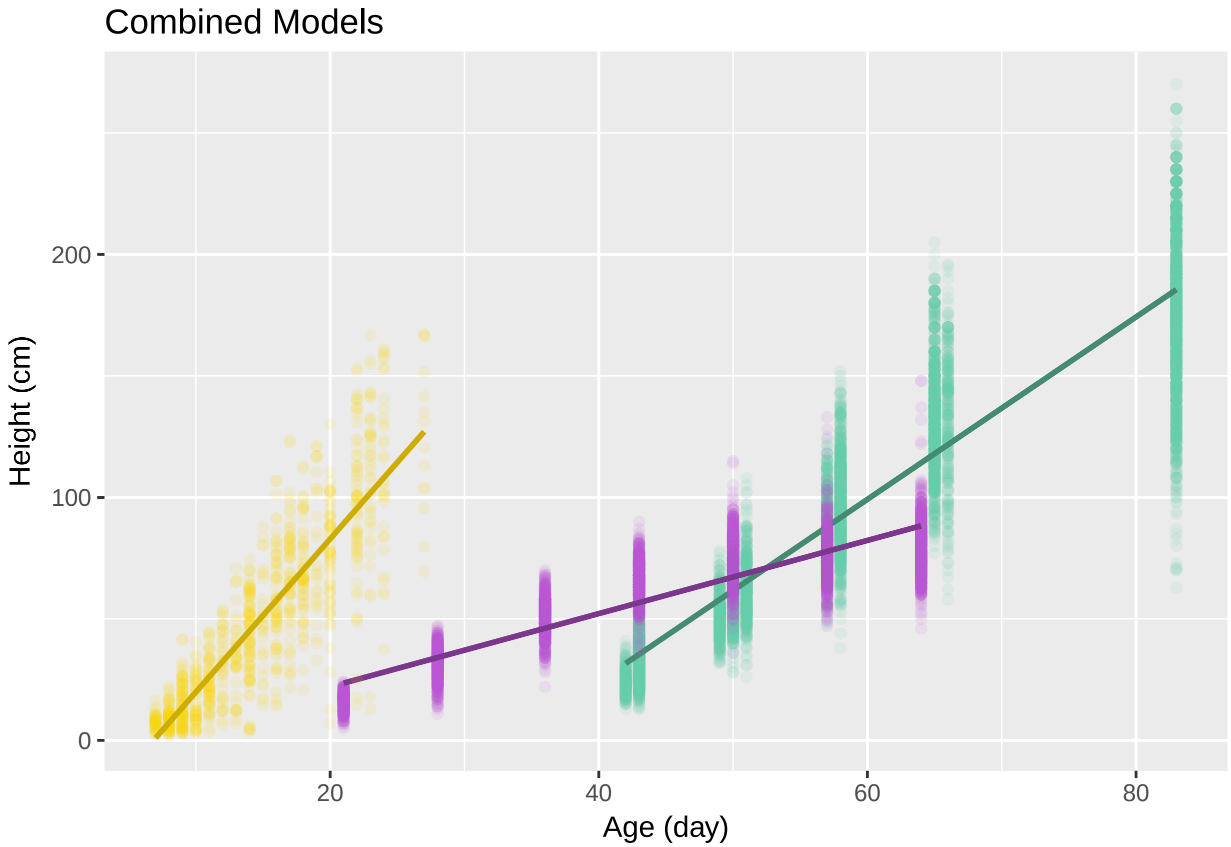
```
ggplot(filter(outdoor_height, season == "Fall 2016"), aes(x = age, y = height_cm)) +
  geom_point(alpha = 1/10) +
  geom_smooth(method = "lm", col = "dodgerblue", se = FALSE) +
  ylab("Height (cm)") +
  xlab("Age (day)") +
  ggtitle("Outdoor - Fall 2016")
```



```
ggsave("outdoor_fall_model.png", path = "~/IndoorOutdoor/data")
```

```
## Saving 6.5 x 4.5 in image
```

```
ggplot() +
  geom_point(data = filter(outdoor_height, season == "Summer 2016"), aes(x = age, y = height_cm), col =
  geom_smooth(data = filter(outdoor_height, season == "Summer 2016"), aes(x = age, y = height_cm), meth
  geom_point(data = filter(outdoor_height, season == "Fall 2016"), aes(x = age, y = height_cm), col = "r
  geom_smooth(data = filter(outdoor_height, season == "Fall 2016"), aes(x = age, y = height_cm), method
  geom_point(data = indoor_height, aes(x = age, y = height_cm), col = "gold", alpha = 1/10) +
  geom_smooth(data = indoor_height, aes(x = age, y = height_cm), method = "lm", col = "gold3", se = FALS
  ylab("Height (cm)") +
  xlab("Age (day)") +
  ggtitle('Combined Models')
```



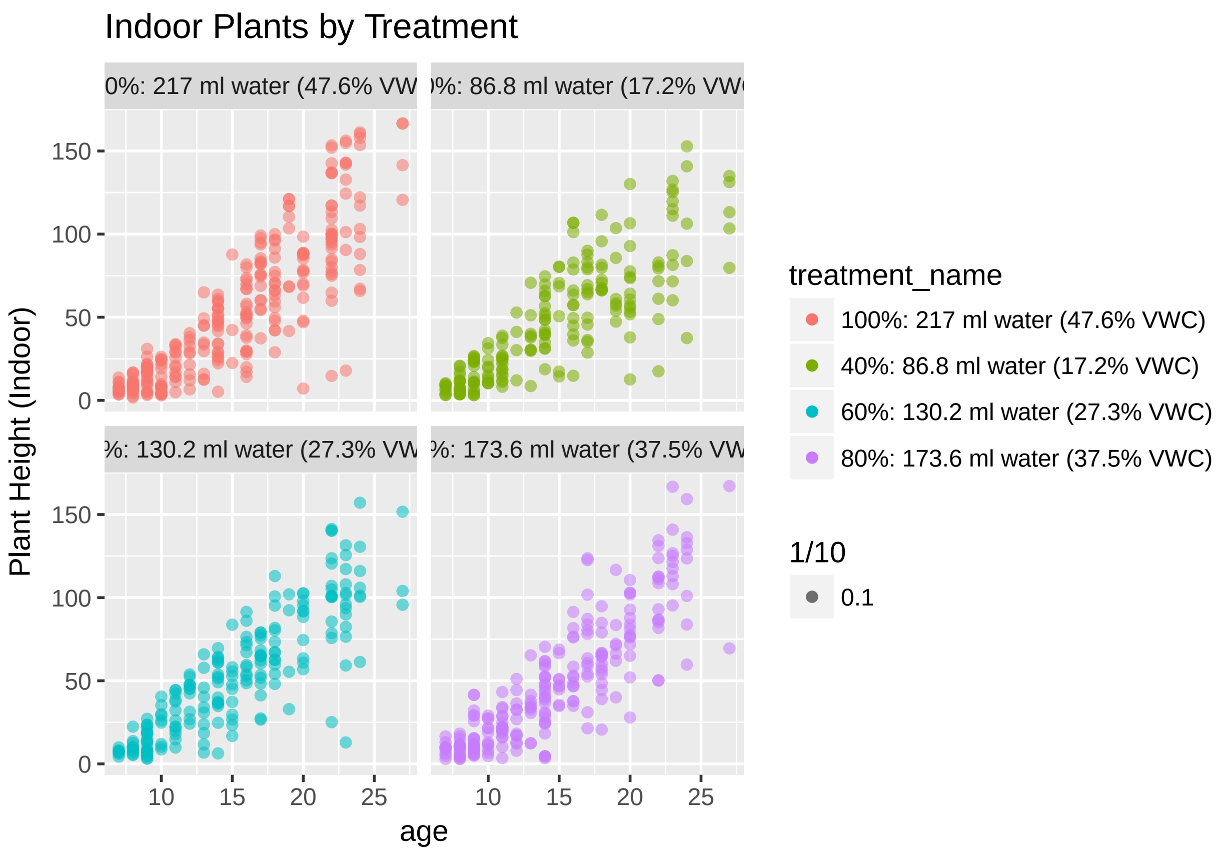
```
ggsave("combined_model.png", path = "~/IndoorOutdoor/data")
```

```
## Saving 6.5 x 4.5 in image
```

Modeling - Yuji

Plots

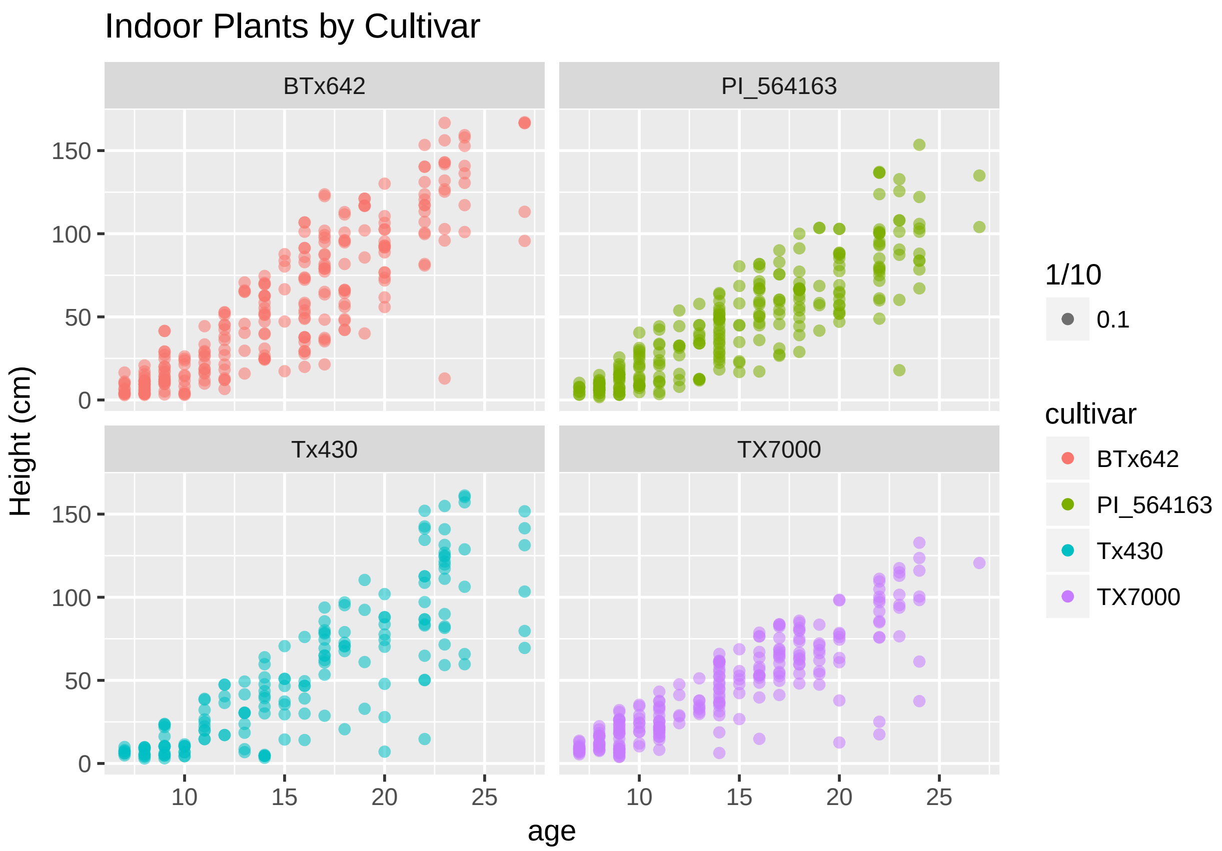
```
## indoor models
ggplot(indoor_height, aes(x = age, y = height_cm)) +
  geom_point(aes(color = treatment_name, alpha = 1/10)) +
  ylab("Plant Height (Indoor)") +
  facet_wrap(~treatment_name) +
  ggtitle("Indoor Plants by Treatment")
```



```
m_indoor_treat <- lm(height_cm ~ age*treatment_name, indoor_height)
summary(m_indoor_treat)
```

```
##
## Call:
## lm(formula = height_cm ~ age * treatment_name, data = indoor_height)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -88.883 -10.163   0.127  10.351  63.374
##
## Coefficients:
##                                                  Estimate Std. Error
## (Intercept)                                      -50.4510     3.6185
## age                                                6.8363     0.2240
## treatment_name40%: 86.8 ml water (17.2% VWC)      17.2315     5.5329
## treatment_name60%: 130.2 ml water (27.3% VWC)     12.3012     5.5861
## treatment_name80%: 173.6 ml water (37.5% VWC)      3.2743     5.4156
## age:treatment_name40%: 86.8 ml water (17.2% VWC)  -1.3200     0.3504
## age:treatment_name60%: 130.2 ml water (27.3% VWC) -0.8070     0.3484
## age:treatment_name80%: 173.6 ml water (37.5% VWC) -0.2927     0.3434
##                                                  t value Pr(>|t|)
## (Intercept)                                      -13.943  < 2e-16 ***
## age                                               30.518  < 2e-16 ***
## treatment_name40%: 86.8 ml water (17.2% VWC)       3.114 0.001901 **
## treatment_name60%: 130.2 ml water (27.3% VWC)      2.202 0.027908 *
## treatment_name80%: 173.6 ml water (37.5% VWC)       0.605 0.545587
## age:treatment_name40%: 86.8 ml water (17.2% VWC)   -3.767 0.000176 ***
## age:treatment_name60%: 130.2 ml water (27.3% VWC)  -2.316 0.020773 *
## age:treatment_name80%: 173.6 ml water (37.5% VWC)  -0.852 0.394299
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.94 on 907 degrees of freedom
## Multiple R-squared:  0.7345, Adjusted R-squared:  0.7325
## F-statistic: 358.5 on 7 and 907 DF,  p-value: < 2.2e-16
```

```
## no significant difference among treatment, except for a slight reduction for 40%

ggplot(indoor_height, aes(x = age, y = height_cm)) +
  geom_point(aes(color = cultivar, alpha = 1/10)) +
  ylab("Height (cm)") +
  facet_wrap(~cultivar) +
  ggtitle("Indoor Plants by Cultivar")
```

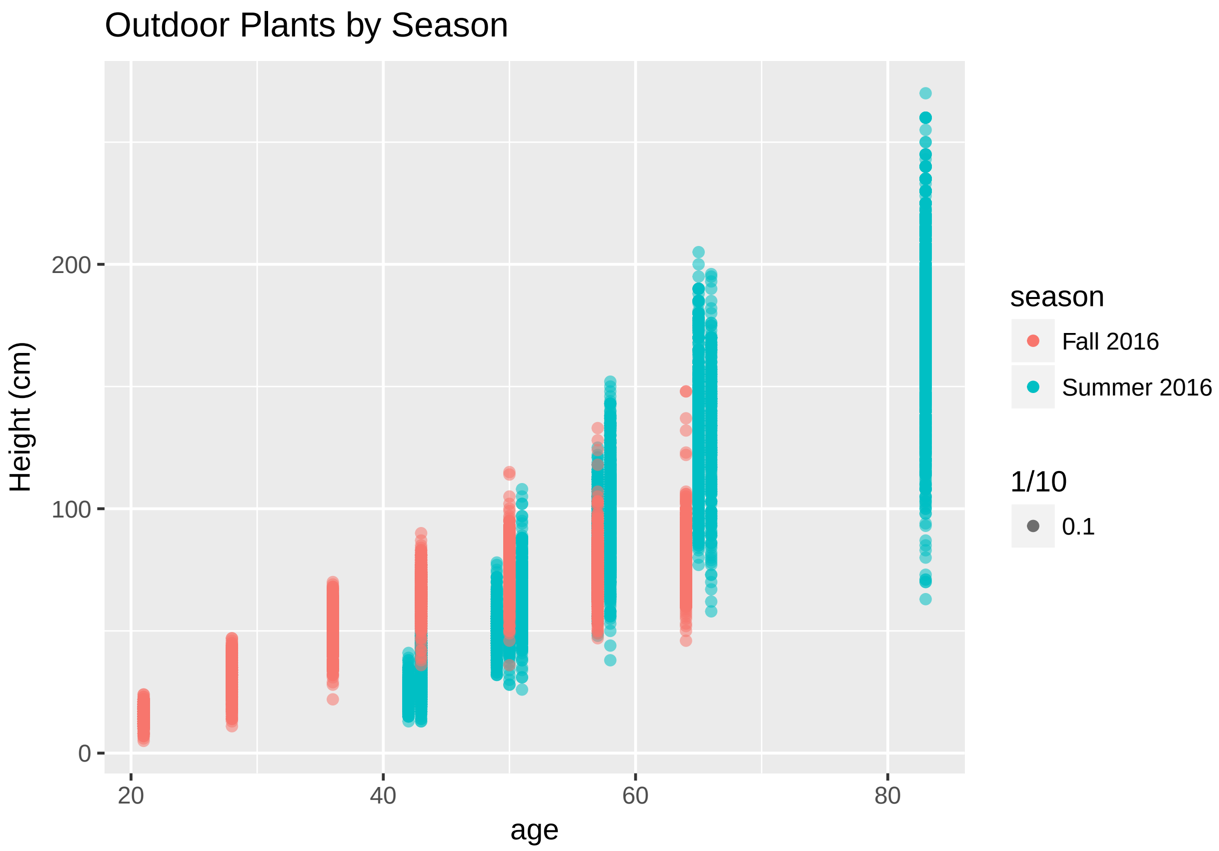


```
m_indoor_cultivar <- lm(height_cm ~ age*cultivar, indoor_height)
summary(m_indoor_cultivar)
```

```
##
## Call:
## lm(formula = height_cm ~ age * cultivar, data = indoor_height)
##
## Residuals:
```

```
##       Min       1Q   Median       3Q      Max 
## -105.532   -9.426    1.100   10.882   58.379 
## 
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)    
## (Intercept)            -53.4956     3.6672 -14.587  < 2e-16 ***
## age                      7.4751     0.2319  32.232  < 2e-16 ***
## cultivarPI_564163       13.5496     5.1884   2.612  0.00916 ** 
## cultivarTx430            4.1634     5.7168   0.728  0.46663    
## cultivarTX7000          22.8443     5.2941   4.315 1.77e-05 ***
## age:cultivarPI_564163   -1.6589     0.3286  -5.049 5.38e-07 ***
## age:cultivarTx430       -0.9773     0.3488  -2.802  0.00519 ** 
## age:cultivarTX7000      -2.1315     0.3441  -6.194 8.90e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 19.14 on 907 degrees of freedom
## Multiple R-squared:  0.7554, Adjusted R-squared:  0.7535 
## F-statistic: 400.1 on 7 and 907 DF,  p-value: < 2.2e-16
```

```
## Growing rates: BTx642 = Tx430 > PI_564163 > TX7000

## outdoor model
ggplot(outdoor_height, aes(x = age, y = height_cm)) +
  geom_point(aes(color = season, alpha = 1/10)) +
  ylab("Height (cm)") +
  ggtitle("Outdoor Plants by Season")
```



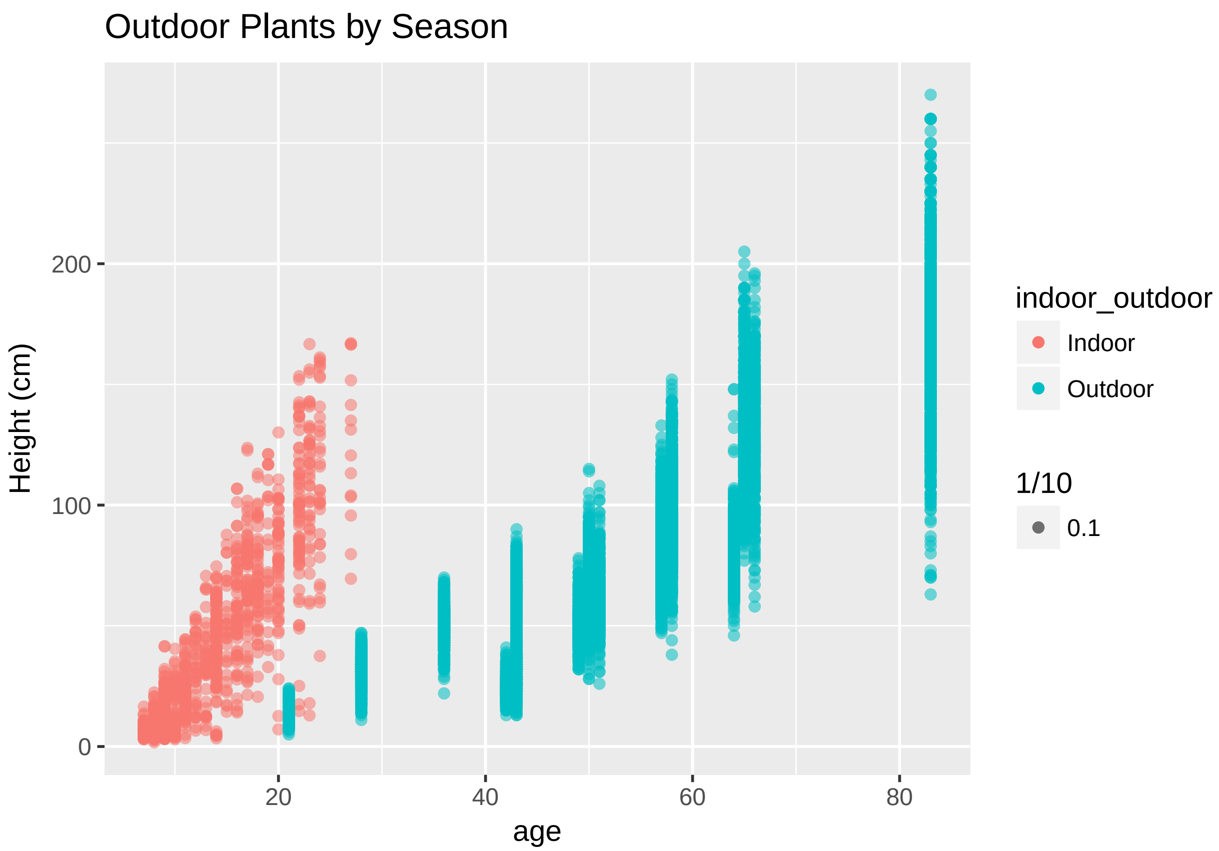
```
m_out_season <- lm(height_cm ~ age*season, data = outdoor_height)
summary(m_out_season)
```

```
## 
## Call:
## lm(formula = height_cm ~ age * season, data = outdoor_height)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -122.525   -8.672   -1.400    8.303   87.031 
## 
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)    
## (Intercept)             -8.07869    0.78635  -10.27   <2e-16 ***
## age                      1.50614    0.01745   86.30   <2e-16 ***
## seasonSummer 2016     -117.90425    1.22173  -96.51   <2e-16 ***
## age:seasonSummer 2016    2.24696    0.02325   96.66   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 17.54 on 11887 degrees of freedom
## Multiple R-squared:  0.8741, Adjusted R-squared:  0.8741 
## F-statistic: 2.751e+04 on 3 and 11887 DF,  p-value: < 2.2e-16
```

```
## growing faster in summer than in fall
```

Put together

```
whole_height <- rbind(indoor_height, outdoor_height)

ggplot(whole_height, aes(x = age, y = height_cm)) +
  geom_point(aes(color = indoor_outdoor, alpha = 1/10)) +
  ylab("Height (cm)") +
  ggtitle("Outdoor Plants by Season")
```



```
## Seems a good model: interaction term accounts for the different growth rates
m_whole <- lm(height_cm ~ age*indoor_outdoor, data = whole_height)
summary(m_whole)
```

```
##
## Call:
## lm(formula = height_cm ~ age * indoor_outdoor, data = whole_height)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -99.001 -18.854   0.633  17.030 107.999
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -43.1856     2.3353 -18.492  < 2e-16 ***
## age                          6.3041     0.1478  42.661  < 2e-16 ***
## indoor_outdoorOutdoor      -17.7409     2.4457  -7.254 4.28e-13 ***
## age:indoor_outdoorOutdoor   -3.6183     0.1484 -24.388  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.33 on 12802 degrees of freedom
## Multiple R-squared:  0.7762, Adjusted R-squared:  0.7761
## F-statistic: 1.48e+04 on 3 and 12802 DF,  p-value: < 2.2e-16
```

```
## Significant difference between indoor and outdoor
```