Data Appendix to "Smith Compost Analysis"

Kate Ginder

Contents

```
library(dummies)
```

```
## dummies-1.5.6 provided by Decision Patterns
```

```
library(caret)
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

1 Appendix description

Your Data Appendix should begin with a brief statement explaining its purpose like the following one.

This Data Appendix documents the data used in "Smith Compost Analysis". It was prepared in a Rmarkdown document that contains both the documentation and the R code used to prepare the data used in the final estimation. It also includes descriptive statistics for both the original data and the final dataset, with a discussion of any issues of note. This data is a time series and will record pounds of compost in a college dinning hall.

The datasets used directly by the final analysis are saved in `processed-data/` at the end of this file.

2 Raw data

This section documents the data sets used in this analysis.

2.1 Dataset description

Citation: Put citation here in APA or other consistent format that you will use throughout the project. Include a hyperlink if applicable.
Date Downloaded: 04/17/2020
Filename: Compost Tracker 3.0.xlsx. **Unit of observation:** amount of compost recorded in dinning halls daily
Dates covered: February 2020 - April 2020

2.1.1 To obtain a copy

To obtain a copy of this data set please contact Susan Sayre at ssayre@smith.edu

2.1.2 Importable version

Filename: importable-data/Raw Data Seminar Paper/Compost Tracker 6.0.csv

The following changes were made to create the importable files.

1.The file was originally opened in excel on a Mac 2.The header reading "Composting Feb & April" was deleted. It was causing the variable names to import incorrectly. 3. Variable names were edited to allow R to read them 4. The document was then saved as a csv file

2.1.3 Variable descriptions

The following data is from two of the Smith College dinning services

- **dates:** Date of the month.

- **Dayoftheweek:** Day of the week the meal is served on.
- **#ofplatesking:** Number of plates per night used in King dinning hall.
- **lbcompostingking:** Pounds of compost per night King dinning hall.
- **#ofplatescutter:** Number of plates per night used in Cutter dinning hall.
- **lbcompostingcutter:** Pounds of compost per night Cutter dinning hall.
- **Meal_number:** Rotating menu cycle the coordinates to a different number.

2.1.4 Data import code and summary

```r
library(readr)
importable_data <- read_csv("Raw Data Seminar Paper/importable_data.csv")
```

```
## Parsed with column specification:
## cols(
##   dates = col_character(),
##   Dayoftheweek = col_character(),
##   `#ofplatesKing` = col_double(),
##   lbcompostKing = col_double(),
##   `#ofplatescutter` = col_double(),
##   lbcompostCutter = col_double(),
##   `meal number` = col_double()
## )
```

```r
View(importable_data)


Compost_data <- read_csv("Raw Data Seminar Paper/importable_data.csv")  %>%
  rename(king_plates = `#ofplatesKing`,
        king_compost = lbcompostKing,
        cutter_plates = `#ofplatescutter`,
        cutter_compost = lbcompostCutter)  %>%
pivot_longer(contains("_"), names_to = c("house","variable"), names_sep = '_', values_to = "values") %>
  pivot_wider(names_from = "variable", values_from = "values") %>%
  mutate(date_var = as.Date(dates, "%m/%d/%y")) %>%
  mutate(cycle = case_when(date_var <= as.Date("2020-03-01") ~ 1,
date_var <= as.Date("2020-04-04") ~ 2,
date_var <= as.Date("2020-05-03") ~ 3))
```

```
## Parsed with column specification:
## cols(
##   dates = col_character(),
##   Dayoftheweek = col_character(),
##   `#ofplatesKing` = col_double(),
##   lbcompostKing = col_double(),
##   `#ofplatescutter` = col_double(),
##   lbcompostCutter = col_double(),
##   `meal number` = col_double()
## )
```

```r
View(Compost_data)
```

3 Data Processing and Construction

4 Analysis Variables

This section should include a description of all the variables that are used in your final analysis. At the end of the section, you should save all of these variables in the processed_data folder of your repository.

Variables used in the final analysis are

- **date_var:** Date of the month read as a numerical value.

- **house:** House variable, King or Cutter.
- **plates:** Number of plates counted per meal.
- **compost:** Pounds of compost collected per night.
- **cycle:** Classifies the rotating menu into three distinct cycles.
- **poster_date:** Dummy variable for no posters (poster date 0: 2/2 - 4/4) and posters poster date 1: 4/5 - 5/2.

The variable 'date_var' originally came from the variable 'dates.' 'Dates' was read as a categorical variables, and it needed to be recognized as a numerical number. These dates space the length of the experiment. The variable 'house' was created by using the pivot function. The data was originally set up for compost and plates in King house and compost and plates in Cutter. By using the prior function house was able to become with own variable. With the option for the two different houses. 'Plates' record the number of plates used per meal. This figure is used to stand as a proxy for the number of student eating in the dinning hall. 'Compost' is recorded in pounds after each mean. 'Cycle' is derived from the 'meal number' variable, from 1-28 indicated the rotating meals for each cycle of the Smith college menu cycle. There are three separate menu cycles. 'Poster_date' is a dummy variable represent the categorical variable the study is looking at. When the posters were not up, the variable is 0. When the posters were up, the variable is 1.

5 Summary Statistics

```
# create dummy variable
Compost_data <- Compost_data %>%
  mutate(poster_date = case_when(date_var <= as.Date("2020-04-04") ~ 0,
date_var <= as.Date("2020-05-2") ~ 1))
```

```
summary(Compost_data)
```

```
dates              Dayoftheweek         meal number        house
```

```
Length:168 Length:168 Min. : 1.00 Length:168
Class :character Class :character 1st Qu.: 7.75 Class :character
Mode :character Mode :character Median :14.50 Mode :character
Mean :14.50
3rd Qu.:21.25
Max. :28.00
plates compost date_var cycle
Min. :121.0 Min. :14.00 Min. :2020-02-02 Min. :1.000
1st Qu.:215.5 1st Qu.:31.44 1st Qu.:2020-02-22 1st Qu.:1.000
Median :334.0 Median :43.38 Median :2020-03-21 Median :2.000
Mean :288.3 Mean :42.54 Mean :2020-03-18 Mean :1.988
3rd Qu.:349.2 3rd Qu.:53.50 3rd Qu.:2020-04-11 3rd Qu.:3.000
Max. :371.0 Max. :78.75 Max. :2020-05-02 Max. :3.000
poster_date
Min. :0.0000
1st Qu.:0.0000
Median :0.0000
```

```
Mean :0.3333
3rd Qu.:1.0000
Max. :1.0000
```

```
king_data <- Compost_data %>% filter(house == "king")

summary(filter(king_data, cycle == 1))
```

```
dates           Dayoftheweek        meal number       house
```

```
Length:29 Length:29 Min. : 1.00 Length:29
Class :character Class :character 1st Qu.: 7.00 Class :character
Mode :character Mode :character Median :14.00 Mode :character
Mean :14.03
3rd Qu.:21.00
Max. :28.00
plates compost date_var cycle poster_date Min. :121 Min. :17.25 Min. :2020-02-02 Min. :1 Min. :0
1st Qu.:210 1st Qu.:32.00 1st Qu.:2020-02-09 1st Qu.:1 1st Qu.:0
Median :333 Median :43.00 Median :2020-02-16 Median :1 Median :0
Mean :285 Mean :41.70 Mean :2020-02-16 Mean :1 Mean :0
3rd Qu.:345 3rd Qu.:52.25 3rd Qu.:2020-02-23 3rd Qu.:1 3rd Qu.:0
Max. :370 Max. :64.25 Max. :2020-03-01 Max. :1 Max. :0
```

```
summary(filter(king_data, cycle == 3))
```

```
dates           Dayoftheweek        meal number       house
```

```
Length:28 Length:28 Min. : 1.00 Length:28
Class :character Class :character 1st Qu.: 7.75 Class :character
Mode :character Mode :character Median :14.50 Mode :character
Mean :14.50
3rd Qu.:21.25
Max. :28.00
plates compost date_var cycle poster_date Min. :138.0 Min. :14.00 Min. :2020-04-05 Min. :3 Min. :1
1st Qu.:230.2 1st Qu.:29.12 1st Qu.:2020-04-11 1st Qu.:3 1st Qu.:1
Median :334.5 Median :40.75 Median :2020-04-18 Median :3 Median :1
Mean :291.7 Mean :40.72 Mean :2020-04-18 Mean :3 Mean :1
3rd Qu.:351.0 3rd Qu.:50.81 3rd Qu.:2020-04-25 3rd Qu.:3 3rd Qu.:1
Max. :368.0 Max. :71.25 Max. :2020-05-02 Max. :3 Max. :1
```

```
sd(filter(king_data, cycle == 1 )$plates)
```

```
[1] 79.4029
```

```
sd(filter(king_data, cycle == 2 )$plates)
```

```
[1] 82.8208
```

```
sd(filter(king_data, cycle == 1 )$compost)
```

```
[1] 12.50828
```

```r
sd(filter(king_data, cycle == 2 )$compost)
```

```
[1] 15.48376
```

```r
cutter_data <- Compost_data %>% filter(house == "cutter")

summary(filter(cutter_data, cycle == 1))
```

```
dates           Dayoftheweek       meal number      house
```

```
Length:29 Length:29 Min. : 1.00 Length:29
Class :character Class :character 1st Qu.: 7.00 Class :character
Mode :character Mode :character Median :14.00 Mode :character
Mean :14.03
3rd Qu.:21.00
Max. :28.00
plates compost date_var cycle poster_date Min. :133 Min. :17.25 Min. :2020-02-02 Min. :1 Min. :0
1st Qu.:208 1st Qu.:29.25 1st Qu.:2020-02-09 1st Qu.:1 1st Qu.:0
Median :327 Median :46.25 Median :2020-02-16 Median :1 Median :0
Mean :289 Mean :41.74 Mean :2020-02-16 Mean :1 Mean :0
3rd Qu.:353 3rd Qu.:53.50 3rd Qu.:2020-02-23 3rd Qu.:1 3rd Qu.:0
Max. :362 Max. :61.00 Max. :2020-03-01 Max. :1 Max. :0
```

```r
summary(filter(cutter_data, cycle == 3))
```

```
dates           Dayoftheweek       meal number      house
```

```
Length:28 Length:28 Min. : 1.00 Length:28
Class :character Class :character 1st Qu.: 7.75 Class :character
Mode :character Mode :character Median :14.50 Mode :character
Mean :14.50
3rd Qu.:21.25
Max. :28.00
plates compost date_var cycle poster_date Min. :134.0 Min. :21.75 Min. :2020-04-05 Min. :3 Min. :1
1st Qu.:224.2 1st Qu.:35.50 1st Qu.:2020-04-11 1st Qu.:3 1st Qu.:1
Median :335.0 Median :47.75 Median :2020-04-18 Median :3 Median :1
Mean :289.2 Mean :43.65 Mean :2020-04-18 Mean :3 Mean :1
3rd Qu.:348.8 3rd Qu.:53.50 3rd Qu.:2020-04-25 3rd Qu.:3 3rd Qu.:1
Max. :364.0 Max. :59.75 Max. :2020-05-02 Max. :3 Max. :1
```

```r
sd(filter(cutter_data, cycle == 1)$plates)
```

```
[1] 77.97069
```

```r
sd(filter(cutter_data, cycle == 2)$plates)
```

```
[1] 80.01683
```

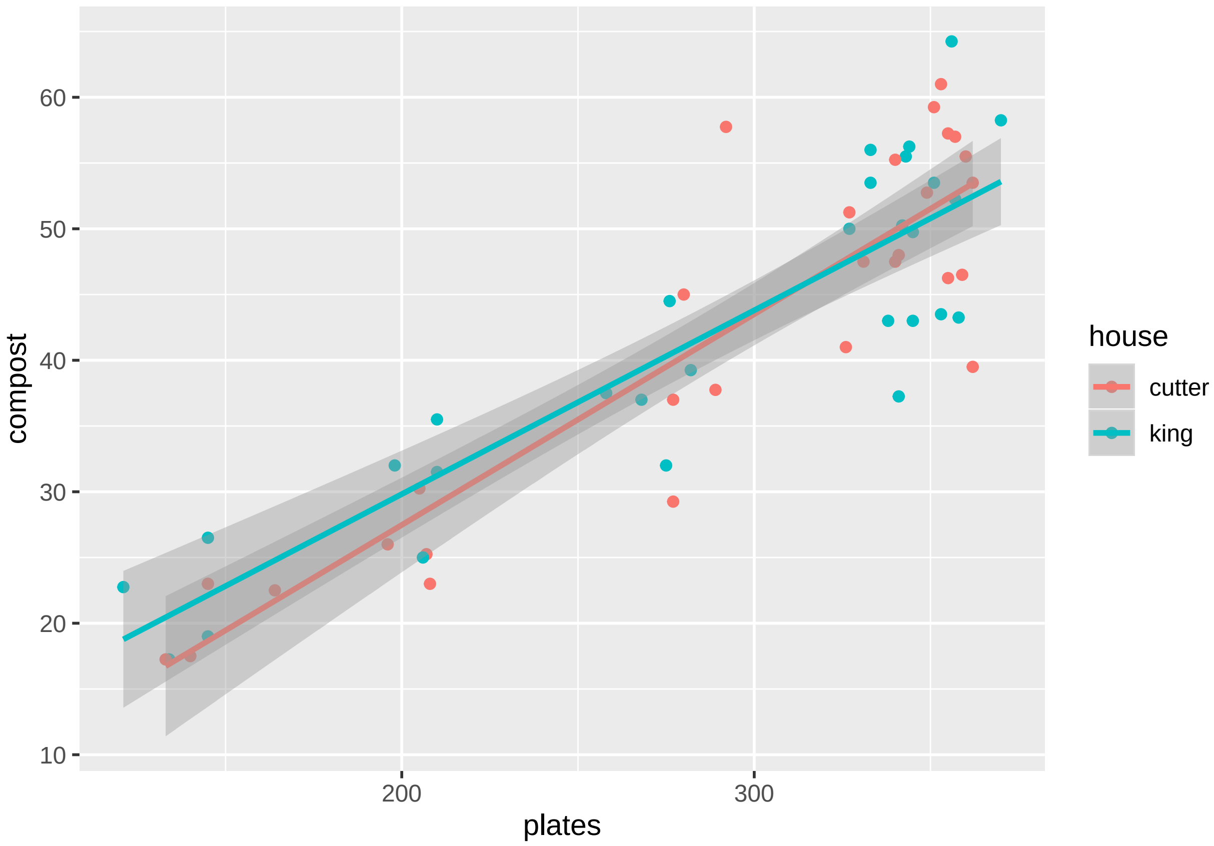
```
sd(filter(cutter_data, cycle == 1)$compost)
```

```
[1] 13.88706
```

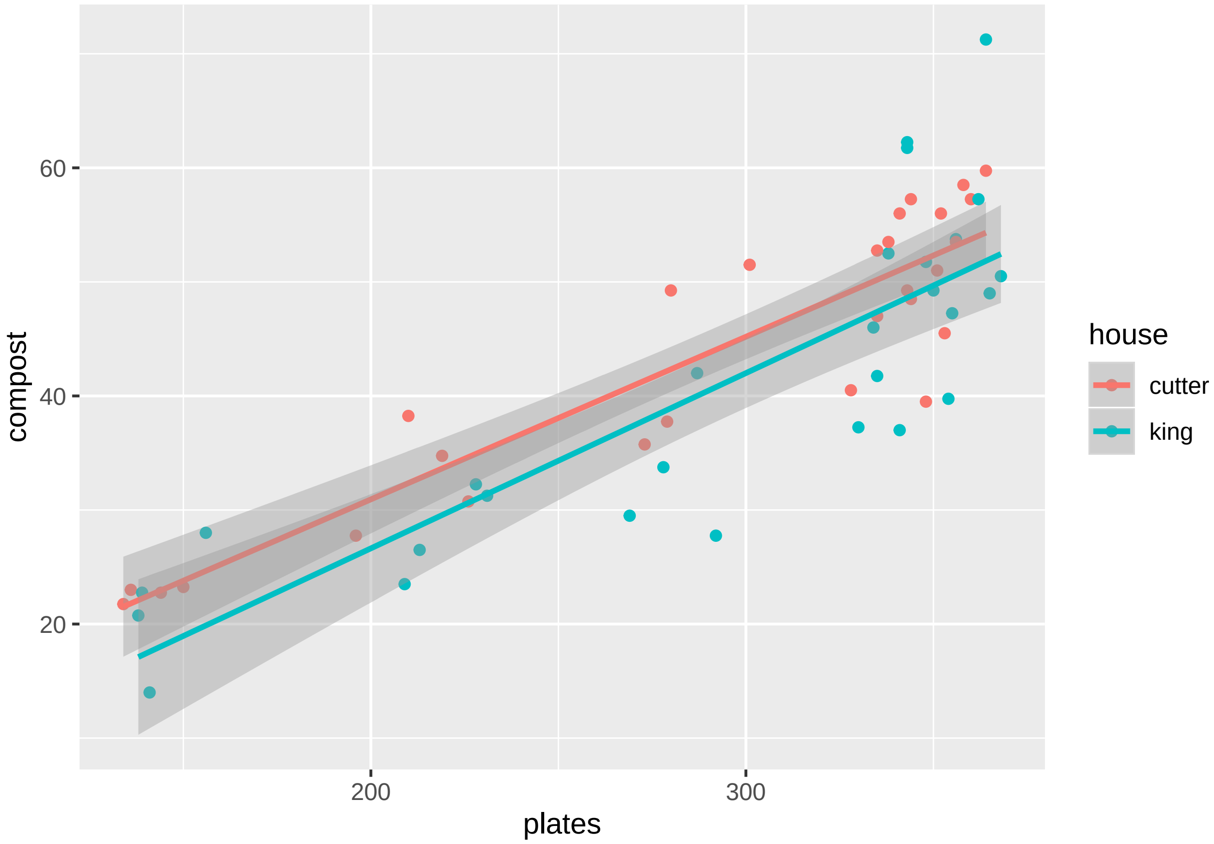
```
sd(filter(cutter_data, cycle == 2)$compost)
```

```
[1] 16.10413
```

```
ggplot(data = filter(Compost_data, cycle == 1), aes(plates, compost, color = house)) + geom_point() + g
```



```
ggplot(data = filter(Compost_data, cycle == 3), aes(plates, compost, color = house)) + geom_point() + g
```

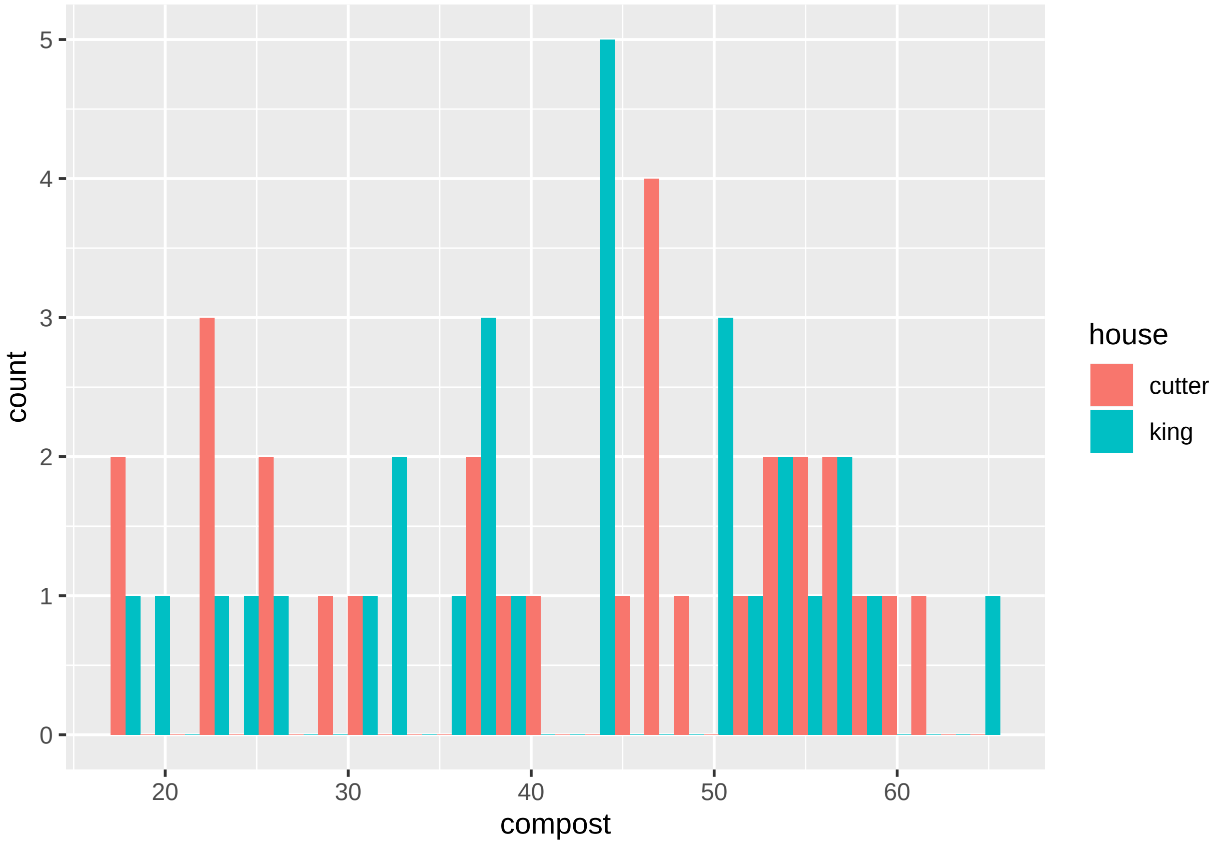


6 Hisograms and Frequency Tables

```
cycle1_results <- Compost_data %>%
  filter(cycle == 1)

ggplot(data = cycle1_results, aes(x = compost, fill = house)) + geom_histogram(position = "dodge")
```

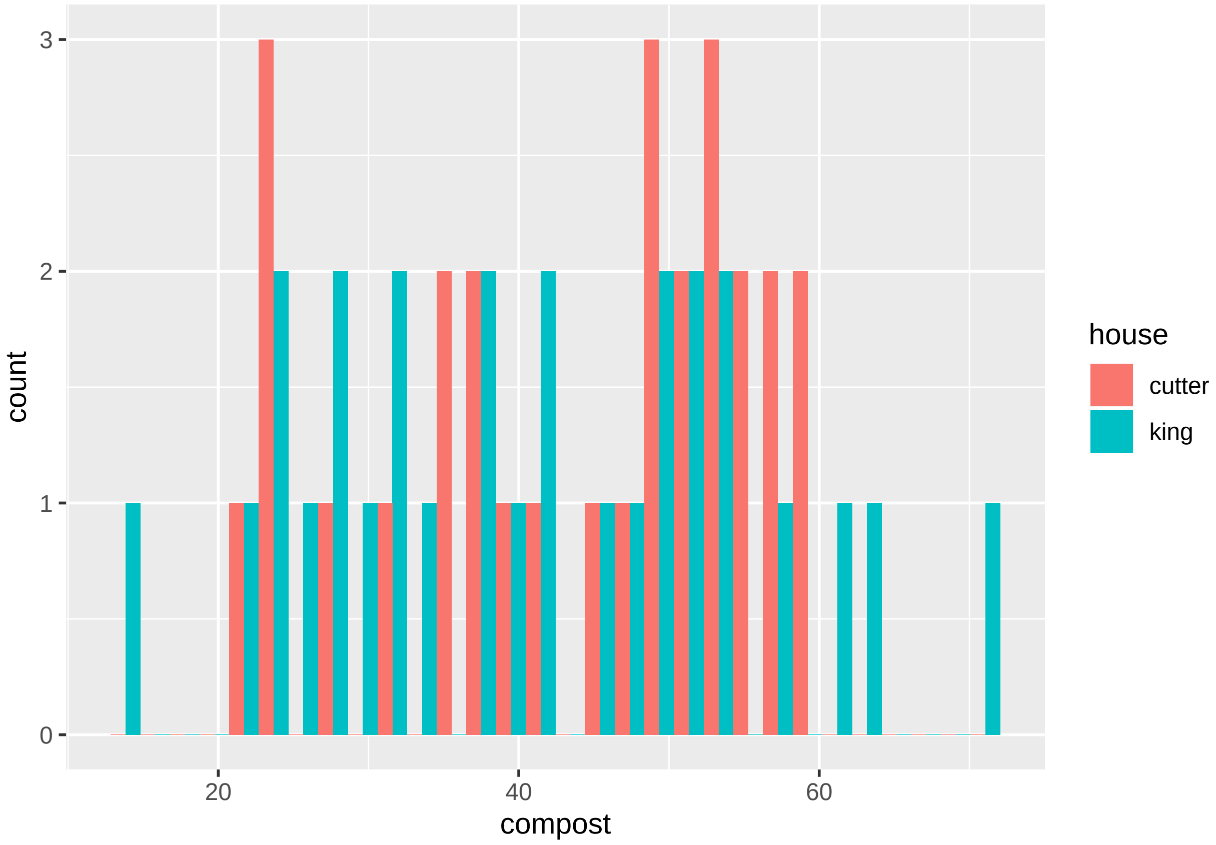
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```r
cycle3_results <- Compost_data %>%
  filter(cycle == 3)

ggplot(data = cycle3_results, aes(x = compost, fill = house)) + geom_histogram(position = "dodge")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



7 Regresssion

```
cor(king_data$compost, king_data$plates)
```

[1] 0.8345402

```
cor(cutter_data$compost, cutter_data$plates)
```

[1] 0.884931

```
linearMod = lm(compost ~ poster_date, data = king_data)
summary(linearMod)
```

Call: lm(formula = compost ~ poster_date, data = king_data)

Residuals: Min 1Q Median 3Q Max -27.772 -11.022 0.353 10.290 34.478

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) 43.022 1.878 22.907 <2e-16 *** poster_date -2.299 3.253 -0.707 0.482
— Signif. codes: 0 ‘’ ***0.001*** ’’ *0.01* ” 0.05 ‘.’ 0.1 ’ ’ 1

Residual standard error: 14.05 on 82 degrees of freedom Multiple R-squared: 0.006054, Adjusted R-squared: -0.006067 F-statistic: 0.4995 on 1 and 82 DF, p-value: 0.4817

```r
linearMod = lm(compost ~ poster_date, data = cutter_data)
summary(linearMod)
```

```
Call: lm(formula = compost ~ poster_date, data = cutter_data)

Residuals: Min 1Q Median 3Q Max -25.165 -12.349 2.217 11.210 36.335

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) 42.415 1.880 22.56 <2e-16 *** poster_date 1.237 3.256 0.38 0.705
— Signif. codes: 0 ‘’ 0.001 ’’ 0.01 ” 0.05 ‘.’ 0.1 ’ ’ 1

Residual standard error: 14.07 on 82 degrees of freedom Multiple R-squared: 0.001756, Adjusted R-squared:
-0.01042 F-statistic: 0.1442 on 1 and 82 DF, p-value: 0.7051
```

```r
stargazer(linearMod, type = "text")
```

```
=============================================== Dependent variable:
——————————— compost
———————————————————— poster_date 1.237
(3.256)

Constant 42.415***
(1.880)

Observations 84
R2 0.002
Adjusted R2 -0.010
Residual Std. Error 14.068 (df = 82)
F Statistic 0.144 (df = 1; 82)
=============================================== Note: p<0.1; p<0.05;
p<0.01
```

8 Discussion of Data

By first looking at the summary statistics for Cutter and King one can see the data is very similar. Because of the size of the two houses, this is to be expected. During the trial period, Cycle 1, King and Cutter have very similar mean pounds of compost 41.70 and 41.74 respectively. They also have similar plate counts 285 and 289 respectively. When the trial begins, Cycle 3, those numbers shift slightly. When King begins the treatment the mean pounds of compost drops to 40.72. Cutter, with no trial implemented, has a mean compost of 43.65 pounds. The mean plates in King during the trial increases slightly to 291.7 while the plates in Cutter moves slightly to 289.2.

The standard deviation for King compost in the baseline period is 12.5 but increases slightly to 15.48 during the trail. This wee can see there in little spread in the overall amount of compost each night. The difference between the 1st and 3rd quartile is at most 20 pounds. Cutter's standard deviation is similar with with a spread of 13.88 during the trial and 16.10 during the trial. For either period the largest difference between the 1st and 3rd quartile is 30 pounds.

A scatter plot of plates and compost shows that there is a general upward trend between the two variables. The more plates used, the more compost is generated. Much of the data is clustered around 340 plates and between 40 and 60 pounds of compost. During cycle 1, the line of best fits are very close. With fewer plates king is producing more compost. The lines intersect at approximately 325 plates and 45 pounds where

Cutter starts producing more compost per plate. In cycle 3, however, the line of best fit for King lies bellow the line for Cutter.

The histograms show the distribution of compost for both cutter and king, but is divided into cycle 1 and cycle 2. The histogram for compost in cycle 1, shows the highest frequencies around 45-50 pounds of compost. The data has a small spread and no outliers. Cycle 3 has higher concentration of frequencies between 20 and 60 pounds of compost compared to cycle 1. The data for Cutter apperears to be slightly bimodle while King's data looks normally distributed.