Data Appendix to "Smith Compost Analysis"

Kate Ginder

Contents


```
library(dplyr)
library(forcats)
```

1 Appendix description

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 dining 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: Smith College Dining Services
Date Downloaded: 04/17/2020
Filename: Compost Tracker 3.0.xlsx. **Unit of observation:** amount of compost recorded in dining 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 King Scales and Cutter Ziskind dining halls at Smith College.

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

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

2.1.4 Data import code and summary

```
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()
## )
```

```
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()
## )
```

```
View(Compost_data)
```

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

3 Data Processing and Construction

```
#reading in the importable data
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()
## )
```

```
View(importable_data)

#renaming the variables into a new BLANk
Compost_data <- read_csv("Raw Data Seminar Paper/importable_data.csv")  %>%
  rename(king_plates = `#ofplatesKing`,
         king_compost = lbcompostKing,
         cutter_plates = `#ofplatescutter`,
         cutter_compost = lbcompostCutter)  %>%
#creating a new variable house
pivot_longer(contains("_"), names_to = c("house","variable"), names_sep = '_', values_to = "values") %>%
  pivot_wider(names_from = "variable", values_from = "values") %>%
#mutating the date variable to be read as a numerical value
  mutate(date_var = as.Date(dates, "%m/%d/%y")) %>%
#creating the Cycle variable
  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()
## )
```

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

#creating a data frame to only include King
king_data <- Compost_data %>% filter(house == "king")

#creating a data frame to only include Cutter
cutter_data <- Compost_data %>% filter(house == "cutter")
```

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 posters (poster date 0: 2/2 - 4/4) (posters poster date 1: 4/5 - 5/2).

The variable 'date_var' originally came from the variable 'dates.' 'Dates' was read as a categorical variable, 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 pivot() function 'house' was able to become its 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 students eating in the dining hall. 'Compost' is recorded in pounds after each mean. 'Cycle' is derived from the 'meal number' variable, from 1-28 indicating the rotating meals for each cycle of the Smith college menu cycle. There are three separate menu cycles. 'Poster_date' is a dummy variable representing the categorical variable the study is looking at. When the posters were not up, the variable was 0. When the posters were up, the variable was 1.

5 Summary Statistics

```
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 :0.1667
3rd Qu.:0.0000
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

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

[1] 15.48376

```
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

```
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. :0
1st Qu.:224.2 1st Qu.:35.50 1st Qu.:2020-04-11 1st Qu.:3 1st Qu.:0
Median :335.0 Median :47.75 Median :2020-04-18 Median :3 Median :0
Mean :289.2 Mean :43.65 Mean :2020-04-18 Mean :3 Mean :0
3rd Qu.:348.8 3rd Qu.:53.50 3rd Qu.:2020-04-25 3rd Qu.:3 3rd Qu.:0
Max. :364.0 Max. :59.75 Max. :2020-05-02 Max. :3 Max. :0

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

[1] 77.97069

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

[1] 80.01683

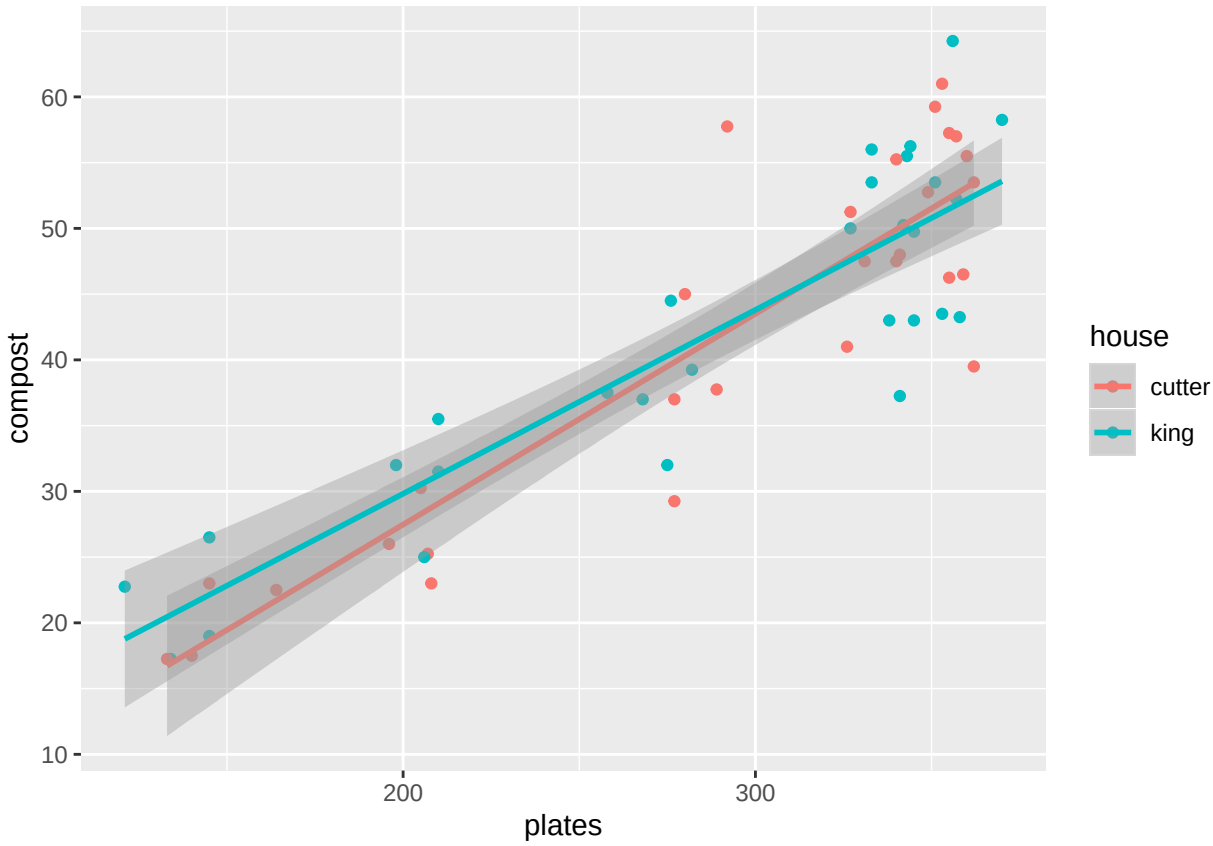
```
sd(filter(cutter_data, cycle == 1)$compost)
```

[1] 13.88706

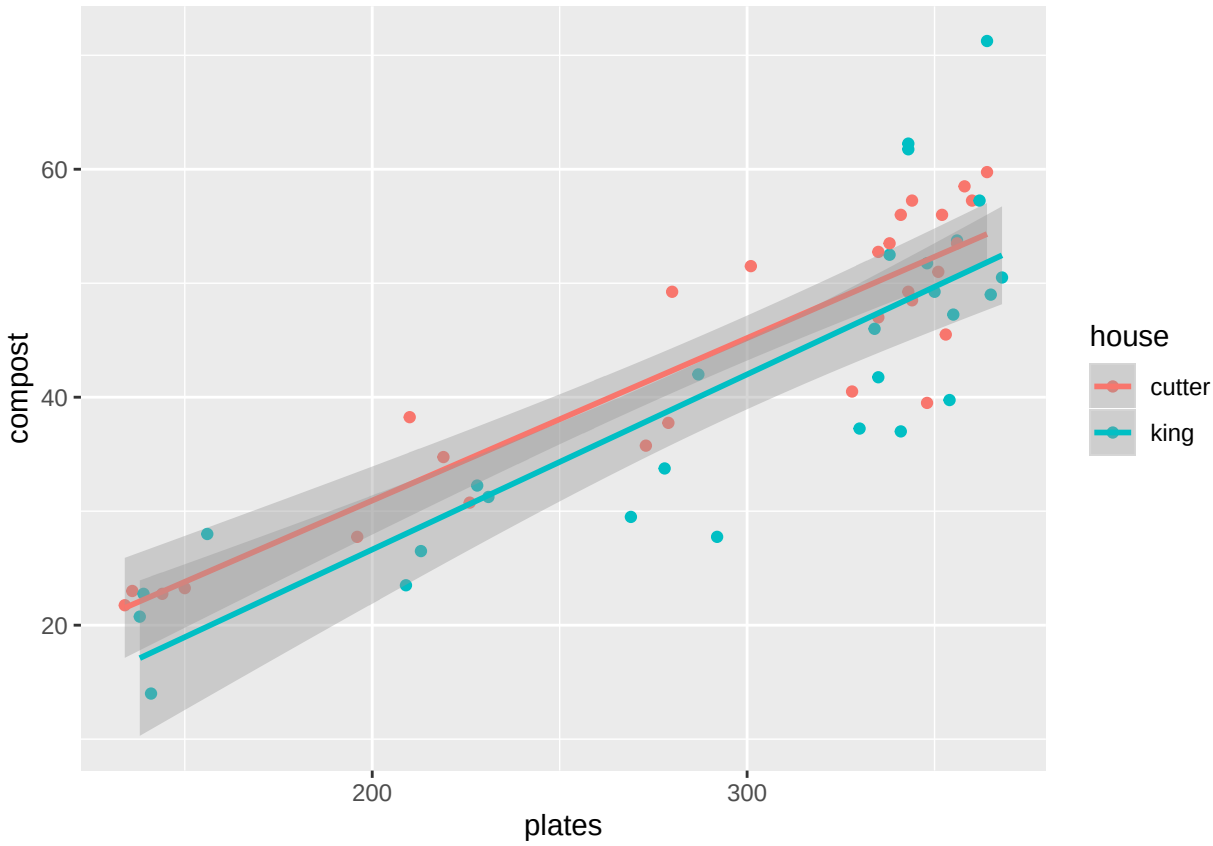
```
sd(filter(cutter_data, cycle == 2)$compost)
```

[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
```

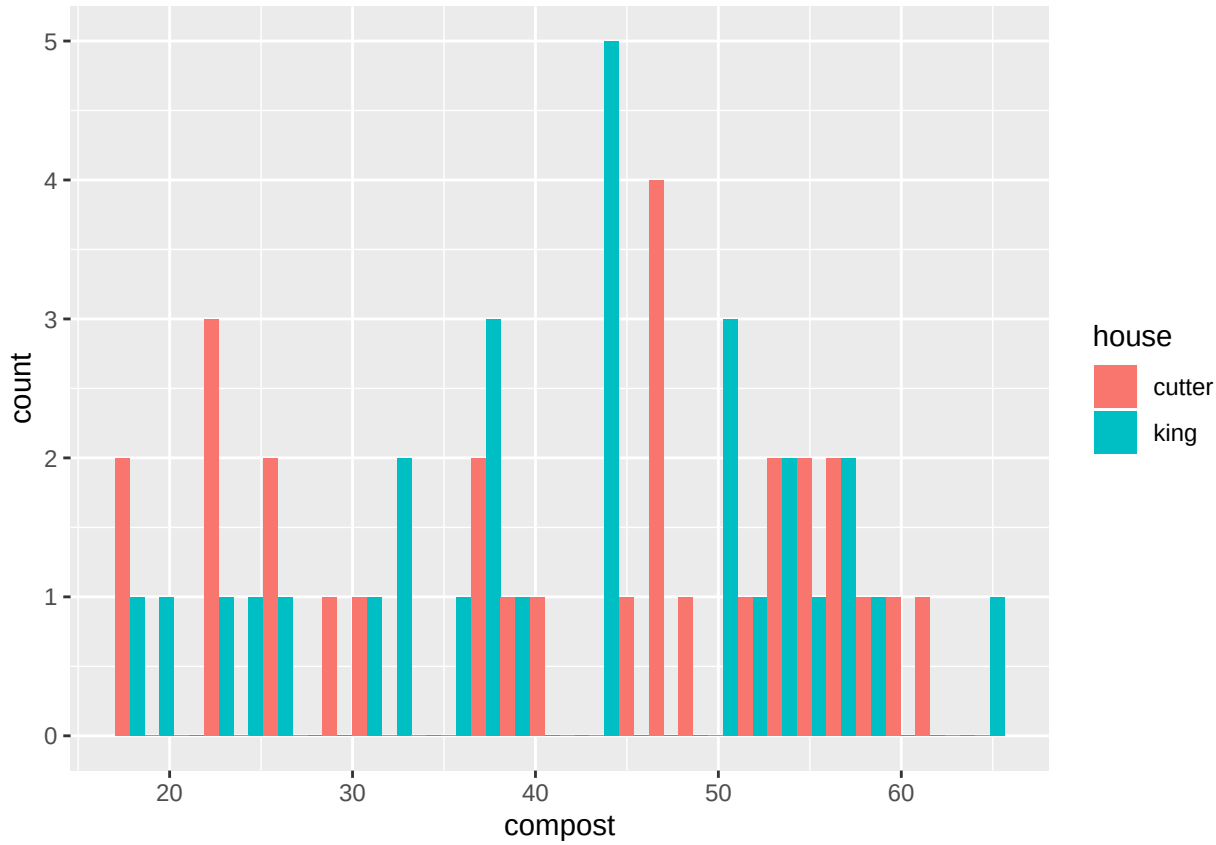


6 Hisograms

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

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

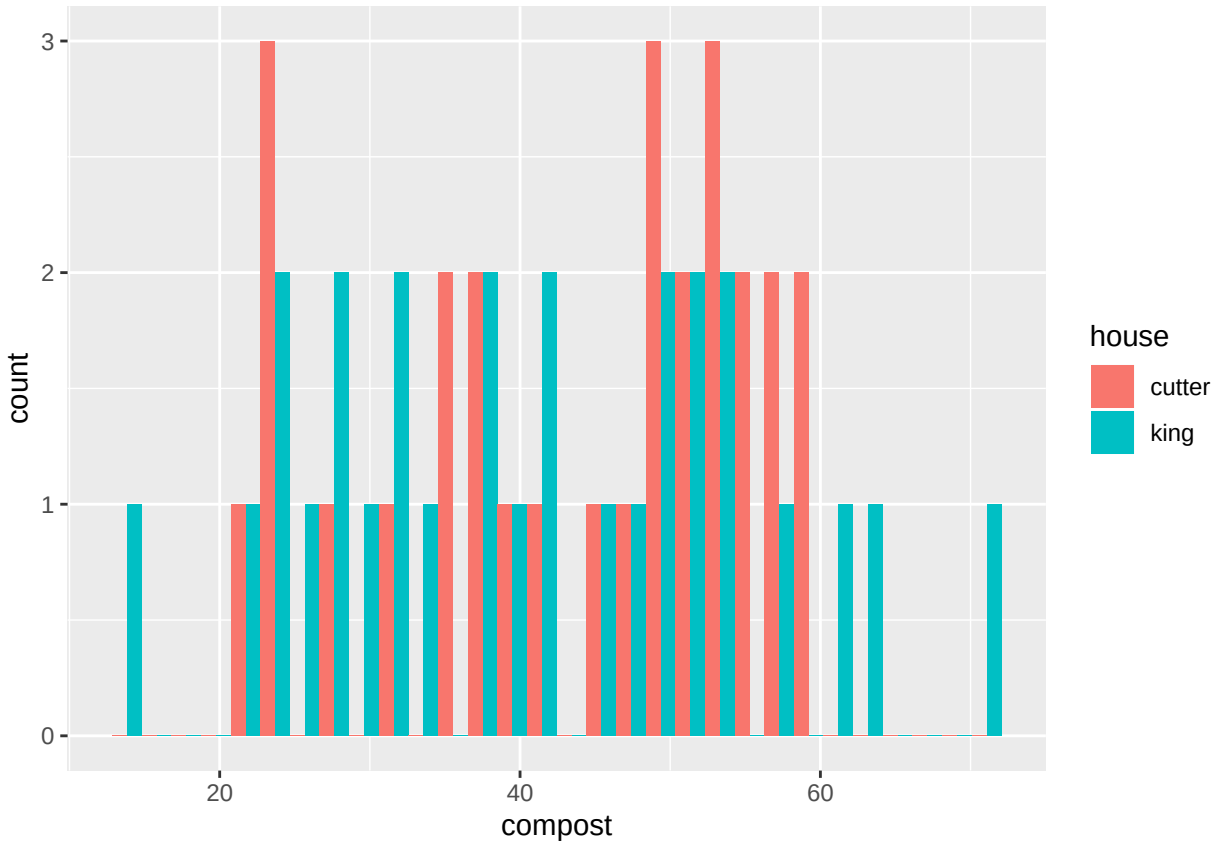
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
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

7.1 Correlation

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

[1] 0.8345402

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

[1] 0.884931

7.2 Regression

```
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

```
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 -25.577 -12.202 2.423 10.798 35.923

Coefficients: (1 not defined because of singularities) Estimate Std. Error t value Pr(>|t|)
(Intercept) 42.827 1.527 28.05 <2e-16 *** poster_date NA NA NA NA
— Signif. codes: 0 ‘’ ***0.001*** ’’ *0.01* ’’ 0.05 ‘.’ 0.1 ’ ’ 1

Residual standard error: 14 on 83 degrees of freedom

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 somewhat to 289.2.

The standard deviation for King compost in the baseline period (cycle 1) is 12.5 lbs but increases slightly to 15.48 lbs during the experiment (cycle 3). Overall, there is minimal spread for total 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 a spread of 13.88 lbs during the baseline (cycle 1) and 16.10 lbs during the trial (cycle 3). 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 lines of best fits for both houses are very close. With fewer plates, King produced more compost. The lines intersect at approximately 325 plates and 45 pounds of compost, where Cutter starts creating more compost per plate. In cycle 3, however, the line of best fit for King lies below the line for Cutter.

The histograms show the distribution of compost for both Cutter and King, but is divided into cycle 1 and cycle 3. 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 a higher concentration of rates between 20 and 60 pounds of compost compared to cycle 1. The data for Cutter appears to be slightly bimodal, while King's data looks normally distributed.

Using there linear model function looking specifically at the King data, we can see that there is a small decrease in compost level with the posters were added. However, due to the late p-value, .482, there is no significance. Additionally, the R^2 value is very small, .006, meaning that the majority of the change in compost level was not due to the poster_date variable.