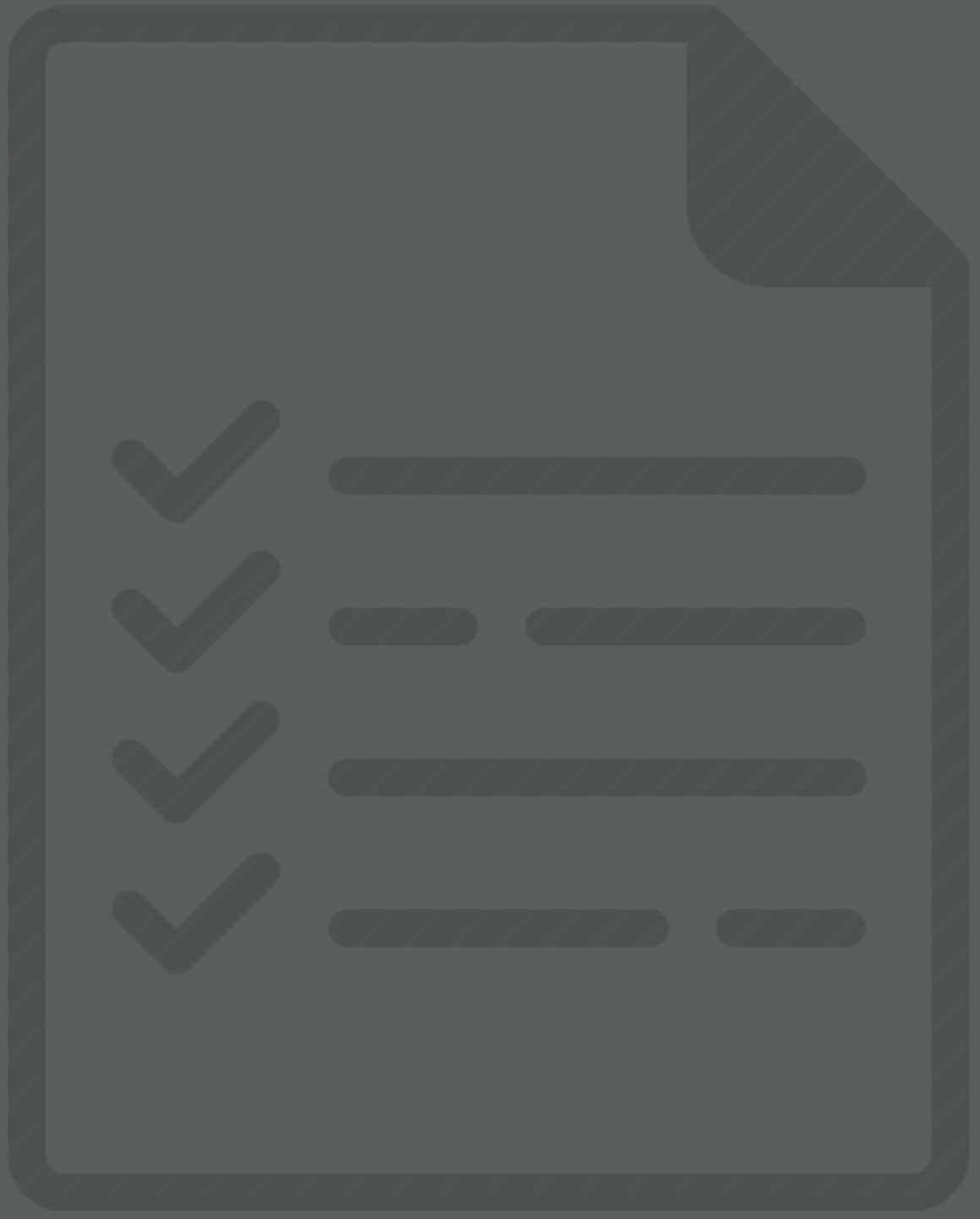
CASE STUDY

Airbnb Frequency Analysis

PREREQUISITES



PREREQUISITES

```
# package prereqs
library(tidyverse)
library(tidytext)
library(magrittr)

# data prereq
airbnb <- read_rds("data/airbnb.rds")
```

Not necessary but I use it for one small function execution.

TASK 1

DESCRIBING AIRBNB LOGDING

1. *Unnest the description text to get single token words*

2. *Remove stop words*

3. *Identify and visualize the most commonly used words to describe*
    - *Apartments*
    - *Houses*

4. *How similar is the word usage between these two property types?*

5. *Can you visualize the 100 words that are used by, and are most common between, the two property types.*

6. *Can you visualize the 100 words that are used by, and are most distinct between, the two property types.*

```r
task1_df <- airbnb %>%
  select(property_type, id, description) %>%
  unnest_tokens(word, description) %>%
  anti_join(stop_words) %>%
  filter(
    property_type %in% c("Apartment", "House"),
    str_detect(word, "[[:alpha:]]")
    ) %>%
  count(property_type, word, sort = TRUE) %>%
  group_by(property_type) %>%
  mutate(pct = n / sum(n))

task1_df
# A tibble: 12,258 x 4
# Groups:   property_type [2]
   property_type word           n     pct
   <chr>         <chr>      <int>   <dbl>
 1 Apartment     boston      3245 0.0190
 2 Apartment     apartment   2995 0.0175
 3 Apartment     kitchen     1951 0.0114
```

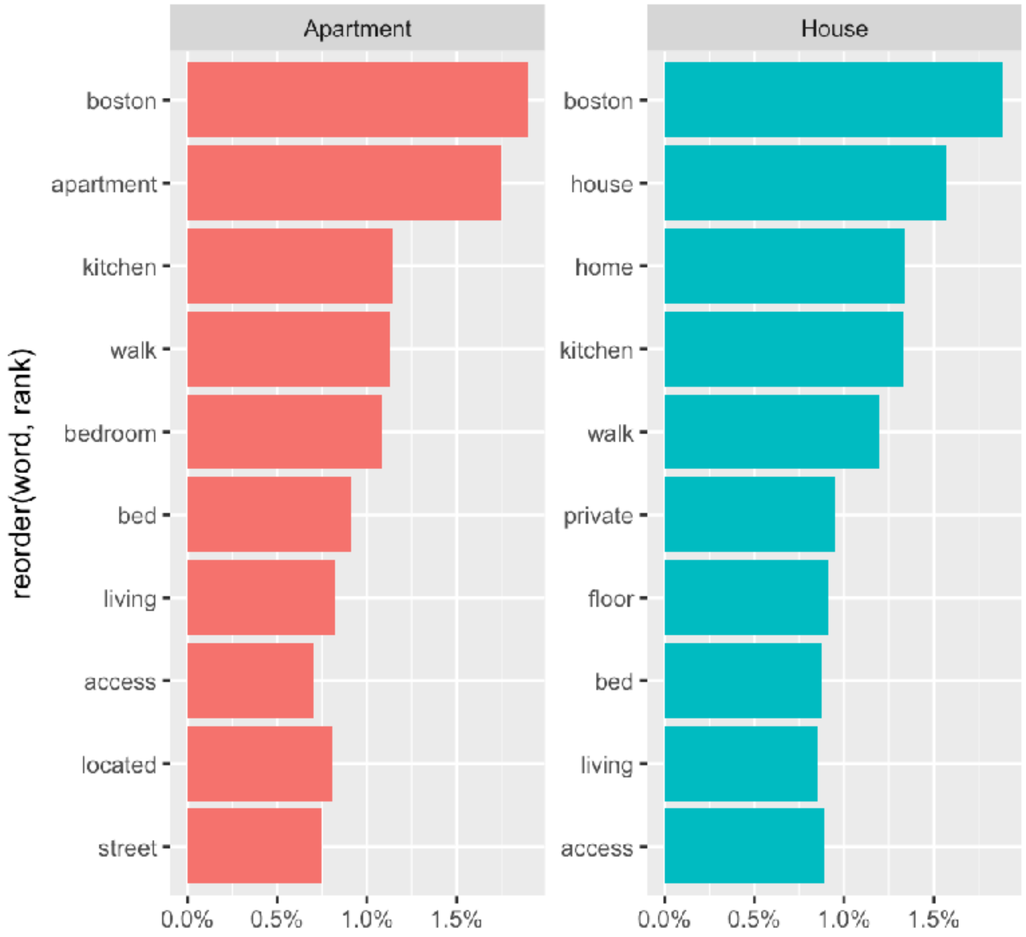
1. Select variables of interest
2. Unnest text to single words
3. Remove stop words

4. Filter for apts & houses
5. Filter out non-alpha words

6. Get word count
7. Compute word proportions by property type

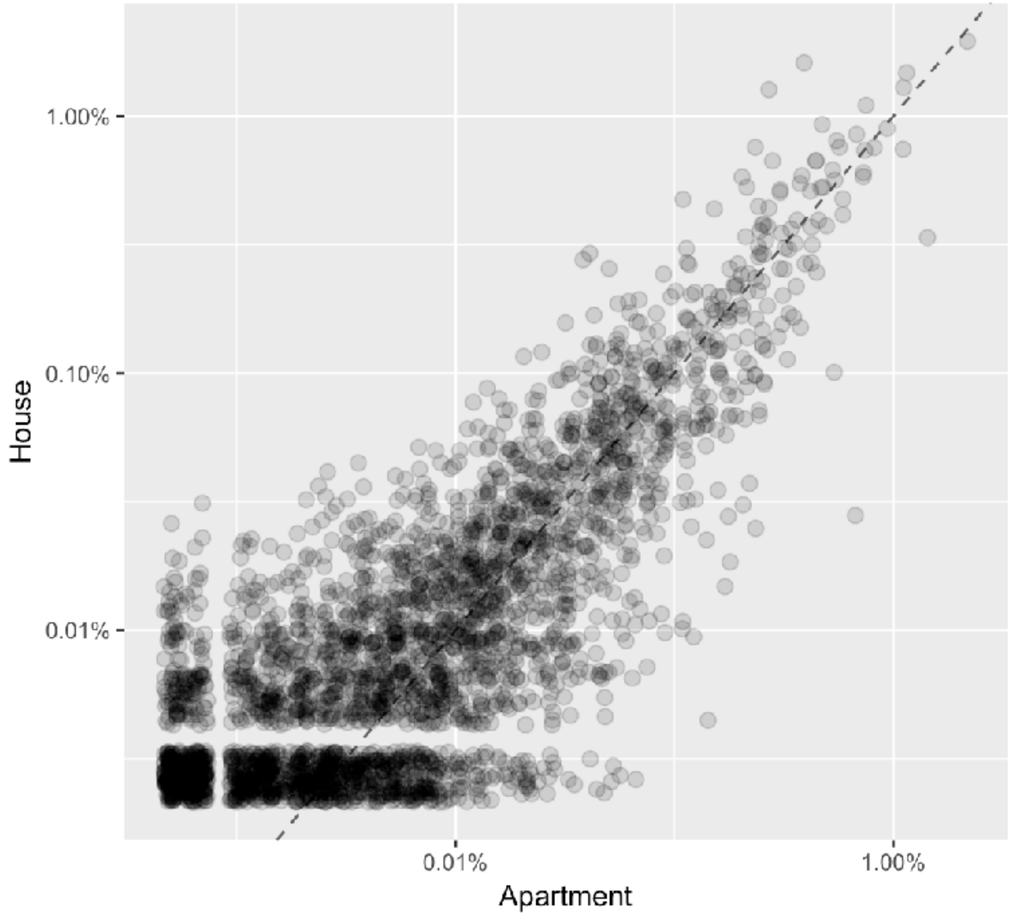
Visualize Top 10 words used to describe apartments and houses

```
task1_df %>%
  top_n(10) %>%
  ungroup() %>%
  arrange(pct) %>%
  mutate(rank = 1:n()) %>%
  ggplot(aes(reorder(word, rank), pct, fill = property_type)) +
  geom_col() +
  scale_y_continuous(NULL, labels = scales::percent) +
  coord_flip() +
  facet_wrap(~ property_type, scales = "free_y")
```



How similar is the word usage between the two property types?

```
task1_df %>%
  select(-n) %>%
  spread(property_type, pct) %>%
  na.omit() %>%
  mutate(delta = abs(Apartment - House)) %>%
  ggplot(aes(Apartment, House)) +
  geom_abline(color = "gray40", lty = 2) +
  geom_jitter(alpha = .15, size = 2.5, width = .1, height = .1) +
  scale_x_log10(labels = scales::percent) +
  scale_y_log10(labels = scales::percent)
```



How similar is the word usage between the two property types?

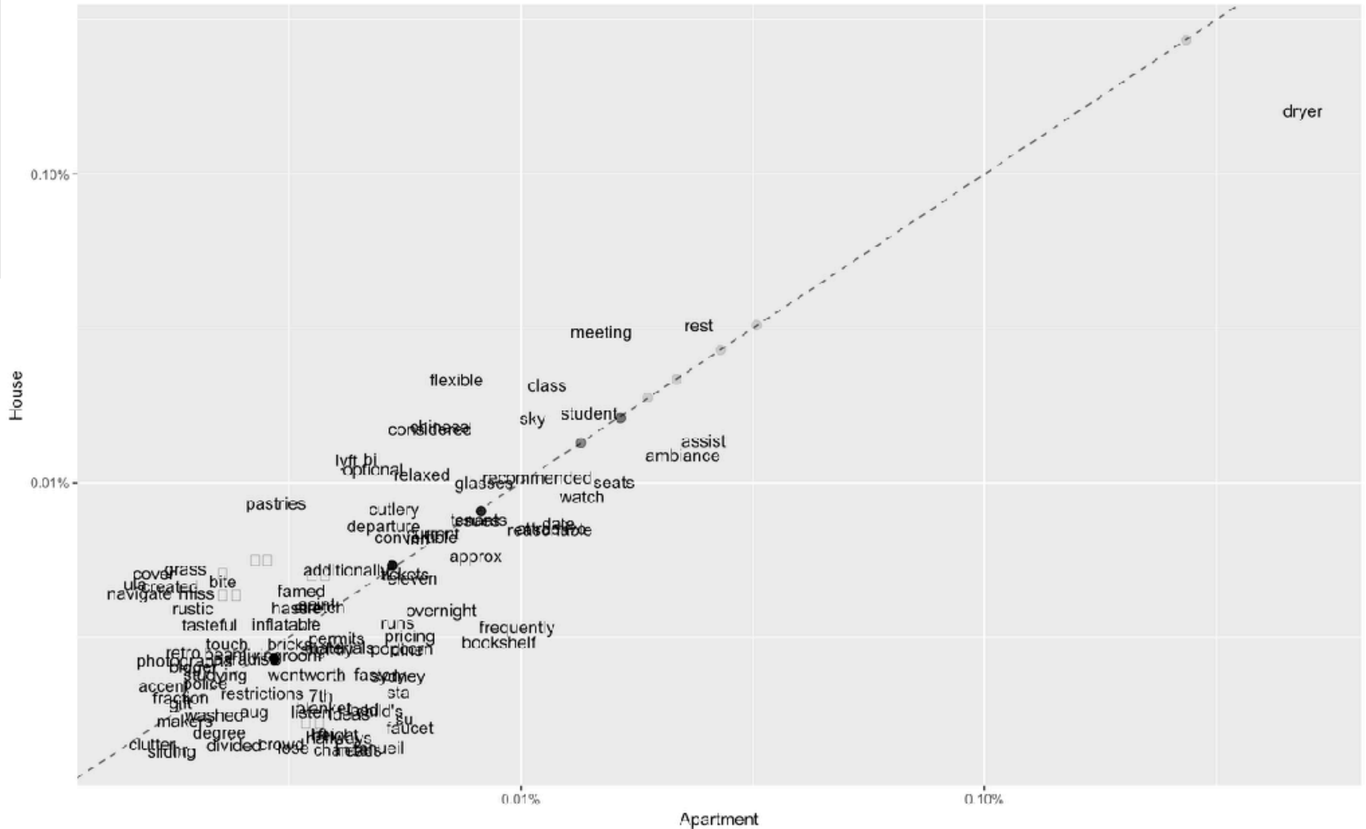
```
task1_df %>%
  select(-n) %>%
  spread(property_type, pct) %>%
  na.omit() %$%
  cor.test(Apartment, House)

 Pearson's product-moment correlation

data:  Apartment and House
t = 87.724, df = 3093, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.834164 0.854384
sample estimates:
      cor
0.8445749
```

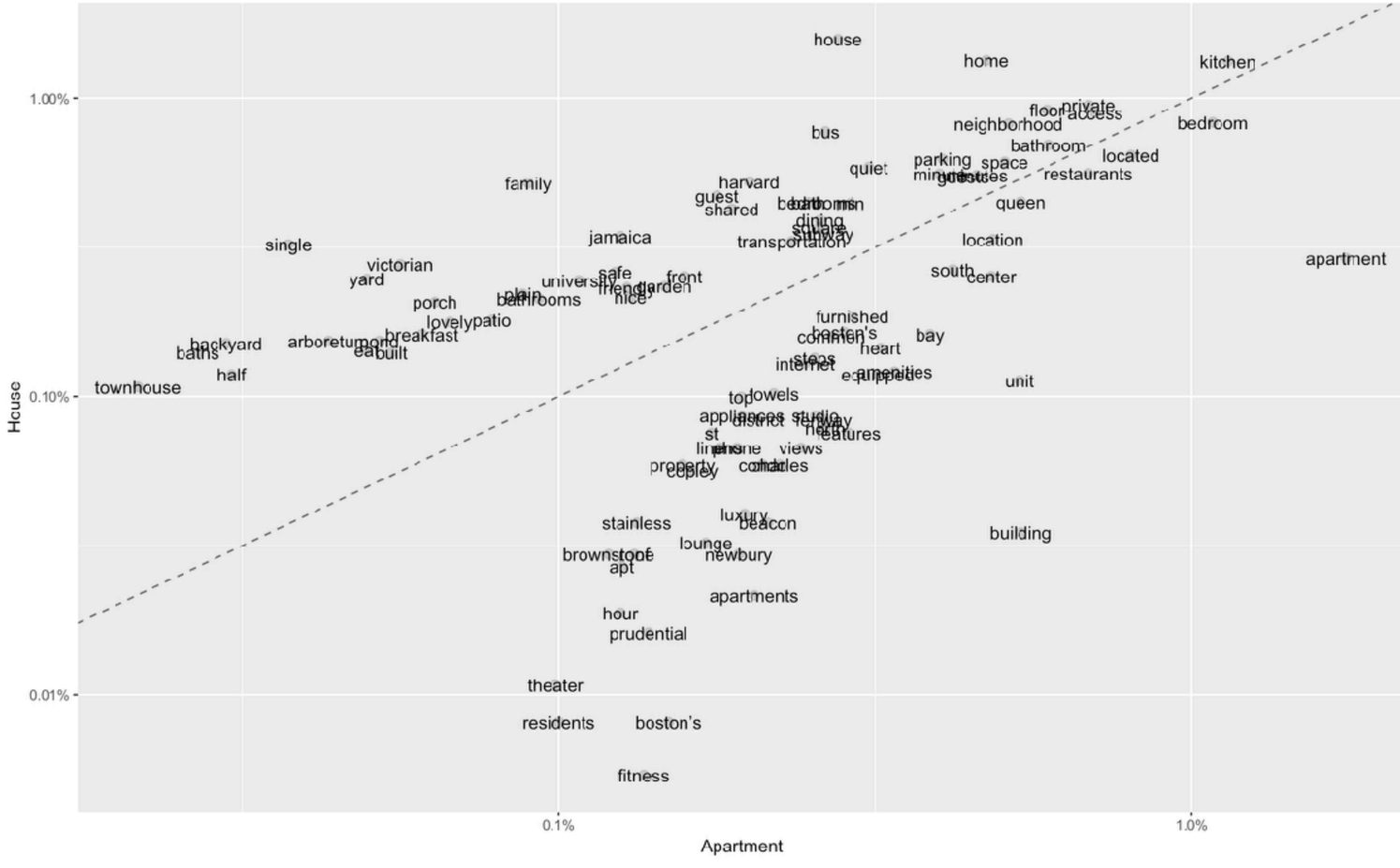
Visualize the 100 words that are used by, and are most common between, the two property types.

```
task1_df %>%
  select(-n) %>%
  spread(property_type, pct) %>%
  na.omit() %>%
  mutate(delta = abs(Apartment - House)) %>%
  top_n(-100, wt = delta) %>%
  ggplot(aes(Apartment, House)) +
  geom_abline(color = "gray40", lty = 2) +
  geom_point(alpha = .15, size = 2.5) +
  geom_text(aes(label = word), position = position_jitter(width = .3, height = .3)) +
  scale_x_log10(labels = scales::percent) +
  scale_y_log10(labels = scales::percent)
```



Visualize the 100 words that are used by, and are most distinct between, the two property types.

```
task1_df %>%
  select(-n) %>%
  spread(property_type, pct) %>%
  na.omit() %>%
  mutate(delta = abs(Apartment - House)) %>%
  top_n(100, wt = delta) %>%
  ggplot(aes(Apartment, House)) +
  geom_abline(color = "gray40", lty = 2) +
  geom_point(alpha = .15, size = 2.5) +
  geom_text(aes(label = word)) +
  scale_x_log10(labels = scales::percent) +
  scale_y_log10(labels = scales::percent)
```



TASK 2

AIRBNB AMENITIES

1. *What are the most commonly provided amenities in the Boston market?*
2. *What proportion of Boston properties are kid friendly?*
3. *Do kid friendly properties tend to get higher or lower prices than those that do not mention kids?*

```
airbnb %>%
  filter(market == "Boston") %>%
  select(id, amenities) %>%
  unnest_tokens(word, amenities) %>%
  anti_join(stop_words) %>%
  count(word, sort = TRUE)
# A tibble: 69 x 2
   word            n
   <chr>       <int>
 1 internet     5975
 2 detector     5325
 3 tv           4361
 4 dryer        4265
 5 wireless     4241
 6 friendly     3694
 7 heating      3367
 8 kitchen      3268
 9 essentials   2977
10 smoke        2895
# ... with 59 more rows
```

1. Filter for Boston properties
2. Select variables of interest
3. Unnest text to words
4. Remove stop words
6. Get word count

What do some of these imply?

Many amenities are explained with multiple words

```
airbnb %>%
  select(amenities) %>%
  filter(str_detect(amenities, "Detector"))
# A tibble: 2,928 x 1
   amenities
   <chr>
 1 "{TV,\"Wireless Internet\",Kitchen,\"Free Parking on Premises\",\"Pets live on this property\",Dog(s),…
 2 "{TV,Internet,\"Wireless Internet\",\"Air Conditioning\",Kitchen,\"Pets Allowed\",\"Pets live on this …
 3 "{TV,\"Cable TV\",\"Wireless Internet\",\"Air Conditioning\",Kitchen,\"Free Parking on Premises\",Heat…
 4 "{TV,Internet,\"Wireless Internet\",\"Air Conditioning\",Kitchen,\"Free Parking on Premises\",Gym,Brea…
 5 "{Internet,\"Wireless Internet\",\"Air Conditioning\",Kitchen,Breakfast,Heating,\"Smoke Detector\",\"C…
 6 "{\"Cable TV\",\"Wireless Internet\",\"Air Conditioning\",Kitchen,\"Free Parking on Premises\",\"Pets …
 7 "{TV,Internet,\"Wireless Internet\",Kitchen,\"Free Parking on Premises\",Heating,\"Smoke Detector\",\"…
 8 "{TV,Internet,\"Wireless Internet\",\"Air Conditioning\",\"Free Parking on Premises\",Breakfast,\"Pets…
 9 "{\"Wireless Internet\",\"Pets live on this property\",Cat(s),Heating,\"Family/Kid Friendly\",\"Smoke …
10 "{TV,\"Cable TV\",Internet,\"Wireless Internet\",\"Air Conditioning\",Kitchen,\"Free Parking on Premis…
# ... with 2,918 more rows
```

One approach: Prior to unnesting we can collapse amenities to a single word

```
airbnb %>%
  filter(market == "Boston") %>%
  select(id, amenities) %>%
  mutate(amenities = str_replace_all(amenities, " ", "")) %>%
  unnest_tokens(word, amenities) %>%
  anti_join(stop_words) %>%
  count(word, sort = TRUE)
# A tibble: 47 x 2
   word                n
   <chr>           <int>
 1 wirelessinternet  3405
 2 heating           3367
 3 kitchen           3268
 4 essentials        2977
 5 smokedetector     2895
```

What proportion of Boston properties are kid friendly?

```
kid_df <- airbnb %>%
  select(price, amenities) %>%
  mutate(kids = str_detect(amenities, regex("kid friendly", ignore_case = TRUE)))

kid_df %>%
  count(kids) %>%
  mutate(pct = n / sum(n))
# A tibble: 2 x 3
  kids      n   pct
  <lgl> <int> <dbl>
1 F      1697 0.473
2 T      1888 0.527
```

Do kid friendly properties tend to get higher or lower prices than those that do not mention kids?

```
kid_price <- kid_df %>%
  mutate(
    price = str_replace_all(price, "(\\$)|(,)|(\\..*)", ""),
    price = str_trim(price) %>% as.numeric()
    )
```

First we need to convert the price variable to a numeric.

This requires removing "$", "," and everything after the ".".

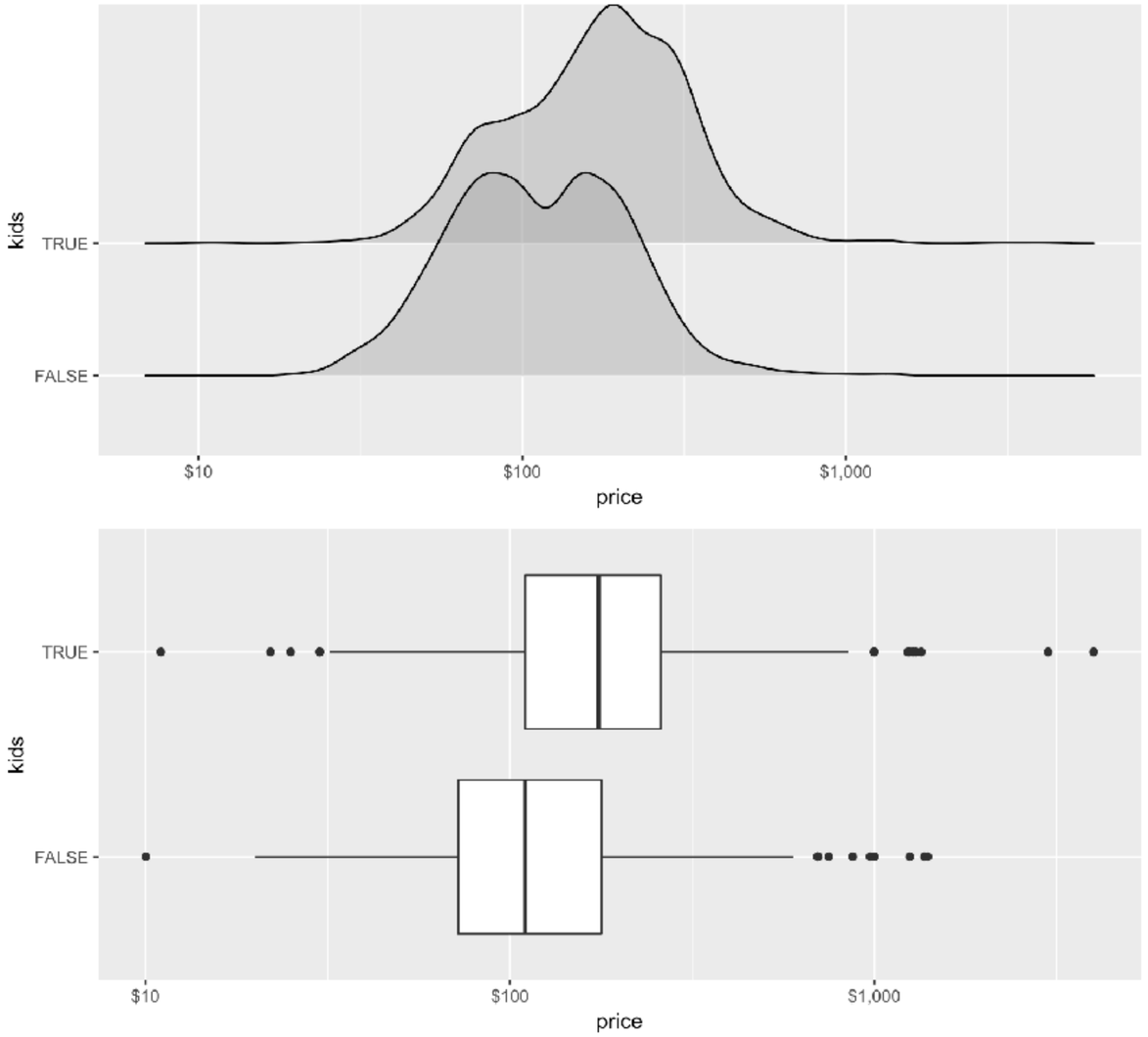
Then trimming any extra white space and converting to a numeric value.

Do kid friendly properties tend to get higher or lower prices than those that do not mention kids?

```
p1 <- ggplot(kid_price, aes(price, kids)) +
  ggridges::geom_density_ridges(alpha = .5) +
  scale_x_log10(labels = scales::dollar)

p2 <- ggplot(kid_price, aes(kids, price)) +
  geom_boxplot() +
  scale_y_log10(labels = scales::dollar) +
  coord_flip()

gridExtra::grid.arrange(p1, p2, ncol = 1)
```



Do kid friendly properties tend to get higher or lower prices than those that do not mention kids?

```
t.test(log(price) ~ kids, data = kid_price)

 Welch Two Sample t-test

data:  log(price) by kids
t = -19.395, df = 3536.3, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.4435931 -0.3621430
sample estimates:
mean in group FALSE  mean in group TRUE
           4.728642            5.131510
```

Yes, there is extremely strong evidence of statistical differences in price

Kid friendly properties charge, on avg, $169/night

Other properties charge, on avg, $113/night