Plotting word counts

INTRODUCTION TO TEXT ANALYSIS IN R

Maham Faisal Khan
Senior Data Science Content Developer

datacamp

Starting with tidy text

```
tidy_review <- review_data %>%
  mutate(id = row_number()) %>%
  unnest_tokens(word, review) %>%
  anti_join(stop_words)
```

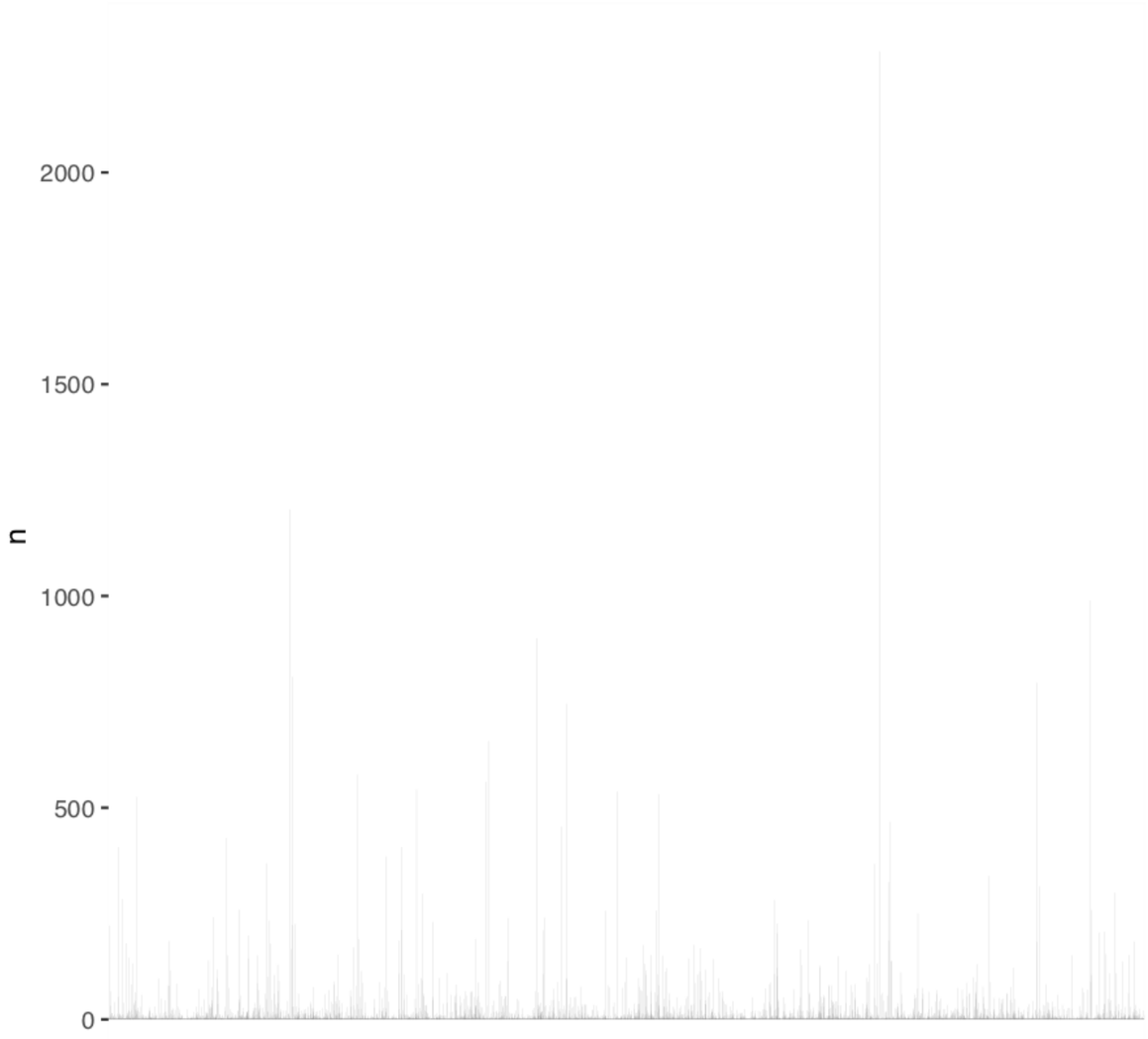
Starting with tidy text

```
tidy_review
```

```
# A tibble: 78,868 x 5
      id date     product                     stars word
   <int> <chr>    <chr>                       <dbl> <chr>
 1     2 1/12/15  iRobot Roomba 650 for Pets      4 walk
 2     2 1/12/15  iRobot Roomba 650 for Pets      4 rest
 3     3 12/26/13 iRobot Roomba 650 for Pets      5 roomba
 4     3 12/26/13 iRobot Roomba 650 for Pets      5 proof
 5     3 12/26/13 iRobot Roomba 650 for Pets      5 house
# … with 78,863 more rows
```

Visualizing counts with geom_col()

```
word_counts <- tidy_review %>%
  count(word) %>%
  arrange(desc(n))
ggplot(
  word_counts, aes(x = word, y = n)
) +
  geom_col()
```



filter() before visualizing

```
word_counts2 <- tidy_review %>%
  count(word) %>%
  filter(n > 300) %>%
  arrange(desc(n))
```

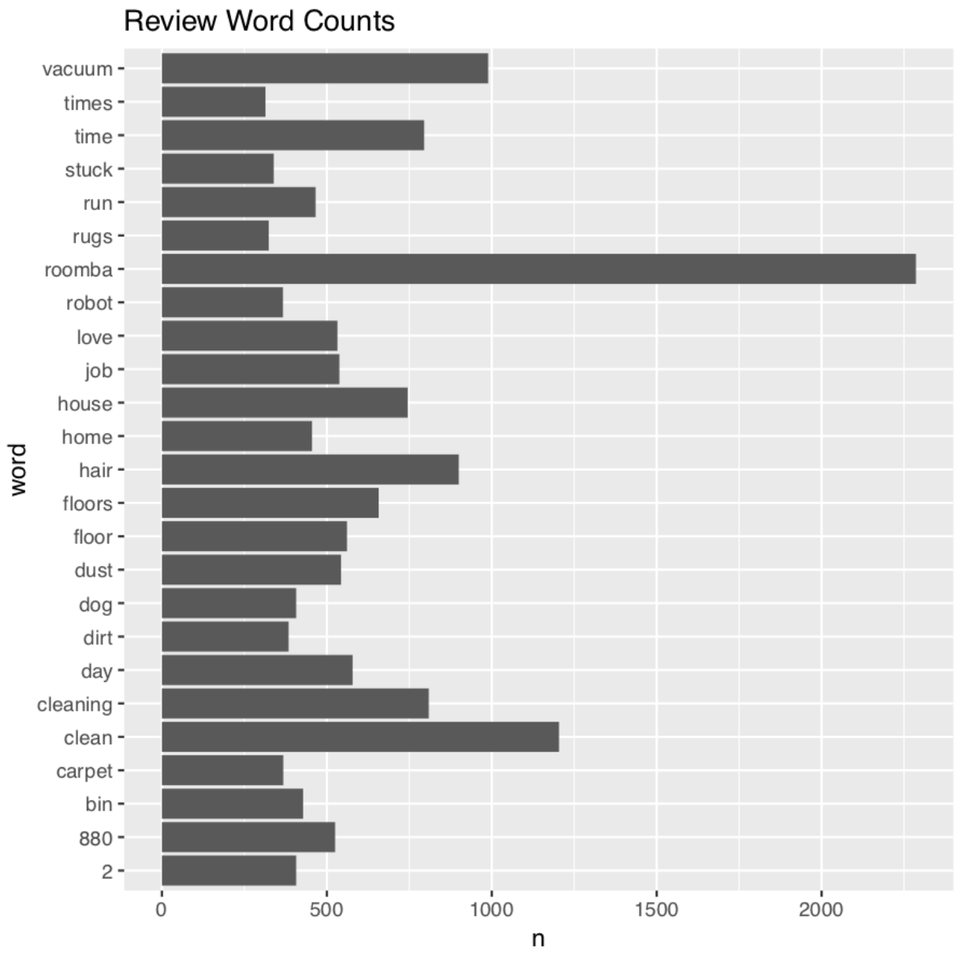
filter() before visualizing

```
word_counts2
```

```
# A tibble: 25 x 2
   word          n
   <chr>     <int>
 1 roomba     2286
 2 clean      1204
 3 vacuum      989
 4 hair        900
 5 cleaning    809
# … with 15 more rows
```

Using coord_flip()

```
ggplot(
  word_counts2, aes(x = word, y = n)
) +
  geom_col() +
  coord_flip() +
  ggtitle("Review Word Counts")
```



Let's practice!

INTRODUCTION TO TEXT ANALYSIS IN R

Improving word count plots

INTRODUCTION TO TEXT ANALYSIS IN R

Maham Faisal Khan
Senior Data Science Content Developer

Custom stop words

```
stop_words
```

```
# A tibble: 1,149 x 2
   word        lexicon
   <chr>       <chr>
 1 a           SMART
 2 a's         SMART
 3 able        SMART
 4 about       SMART
 5 above       SMART
# … with 1,144 more rows
```

Using tribble()

```
tribble(
  ~word,    ~lexicon,
  "roomba", "CUSTOM",
  "2",      "CUSTOM"
)
```

```
# A tibble: 2 x 2
  word   lexicon
  <chr>  <chr>
1 roomba CUSTOM
2 2      CUSTOM
```

Using bind_rows()

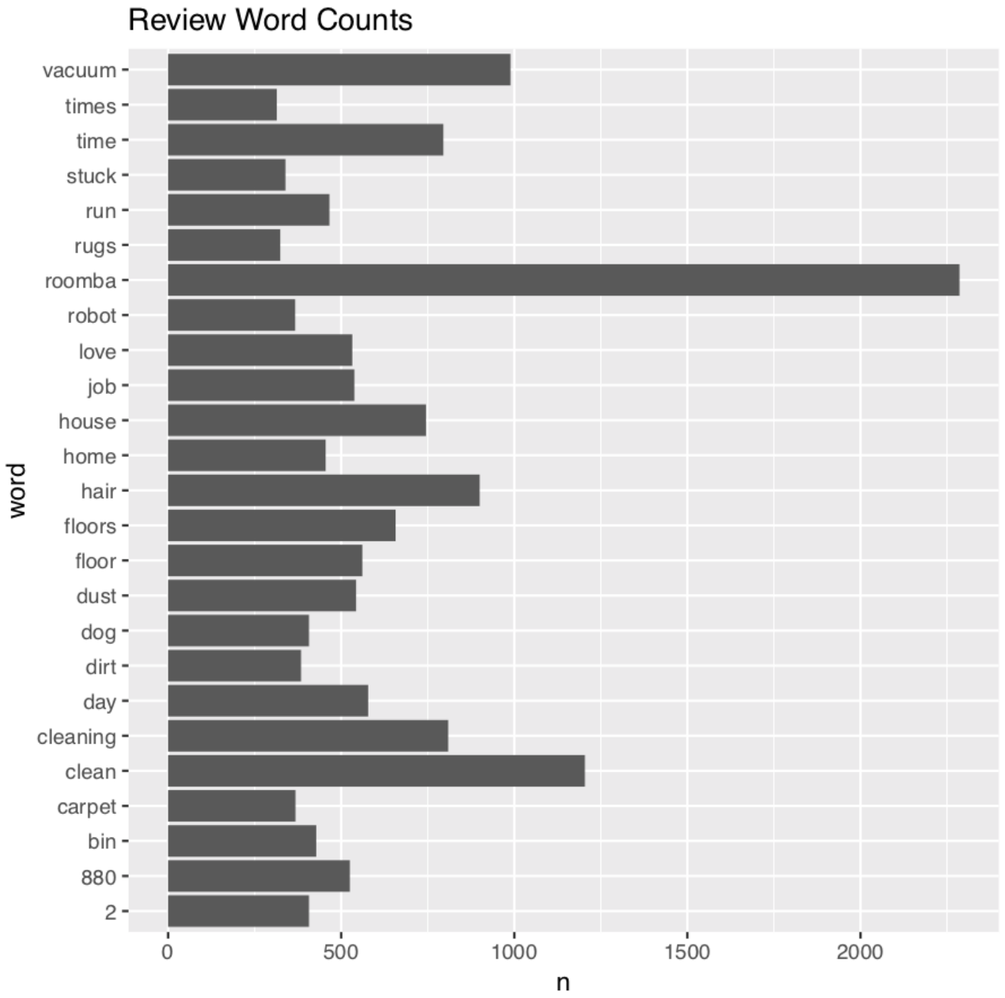
```
custom_stop_words <- tribble(
  ~word,    ~lexicon,
  "roomba", "CUSTOM",
  "2",      "CUSTOM"
)
stop_words2 <- stop_words %>%
  bind_rows(custom_stop_words)
```

Removing stop words again

```
tidy_review <- review_data %>%
  mutate(id = row_number()) %>%
  select(id, date, product, stars, review) %>%
  unnest_tokens(word, review) %>%
  anti_join(stop_words2)
tidy_review %>%
  filter(word == "roomba")
```

```
# A tibble: 0 x 5
# … with 5 variables: id <int>, date <chr>, product <chr>, stars <dbl>, word <chr>
```

Factors


Review Word Counts
vacuum
times
time
stuck
run
rugs
roomba
robot
love
job
house
home
hair
floors
floor
dust
dog
dirt
day
cleaning
clean
carpet
bin
880
2
word
0
500
1000
1500
2000
n

Using fct_reorder()

```
word_counts <- tidy_review %>%
  count(word) %>%
  filter(n > 300) %>%
  mutate(word2 = fct_reorder(word, n))
```

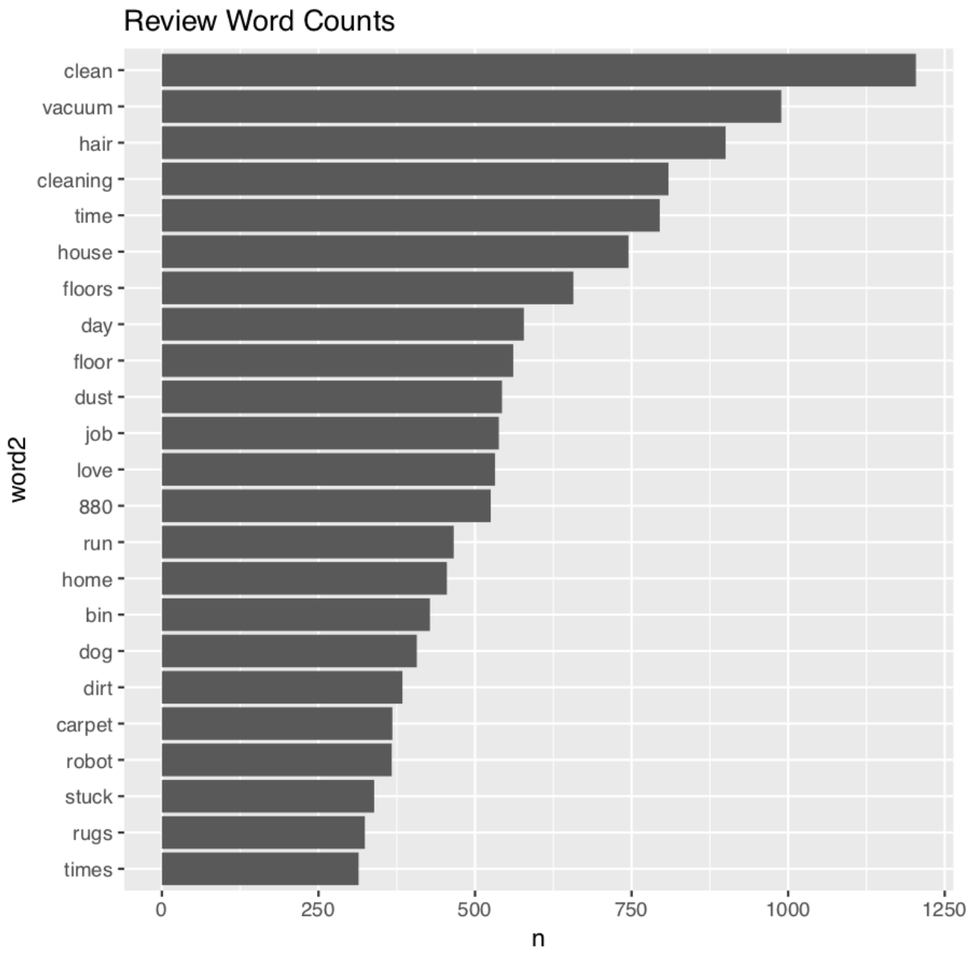
Using fct_reorder()

```
word_counts
```

```
# A tibble: 23 x 3
   word          n word2
   <chr>     <int> <fct>
 1 880         525 880
 2 bin         428 bin
# … with 21 more rows
```

Arranging the bar plot

```
ggplot(
  word_counts, aes(x = word2, y = n)
) +
  geom_col() +
  coord_flip() +
  ggtitle("Review Word Counts")
```



Let's practice!

INTRODUCTION TO TEXT ANALYSIS IN R

Faceting word count plots

INTRODUCTION TO TEXT ANALYSIS IN R

Maham Faisal Khan
Senior Data Science Content Developer

Counting by product

```
tidy_review %>%
  count(word, product) %>%
  arrange(desc(n))
```

```
# A tibble: 12,719 x 3
   word     product                                          n
   <chr>    <chr>                                        <int>
 1 clean    iRobot Roomba 880 for Pets and Allergies   815
 2 vacuum   iRobot Roomba 880 for Pets and Allergies   678
 3 hair     iRobot Roomba 880 for Pets and Allergies   595
# … with 12,716 more rows
```

Using slice_max()

```
tidy_review %>%
  count(word, product) %>%
  group_by(product) %>%
  slice_max(n, n = 10)
```

```
# A tibble: 20 x 3
# Groups:   product [2]
   word     product                                n
   <chr>    <chr>                              <int>
 1 650      iRobot Roomba 650 for Pets           108
# … with 19 more rows
```

Using ungroup()

```
tidy_review %>%
  count(word, product) %>%
  group_by(product) %>%
  slice_max(n, n = 10) %>%
  ungroup()
```

```
# A tibble: 10 x 3
   word     product                                    n
   <chr>    <chr>                                  <int>
 1 650      iRobot Roomba 650 for Pets               108
# … with 9 more rows
```

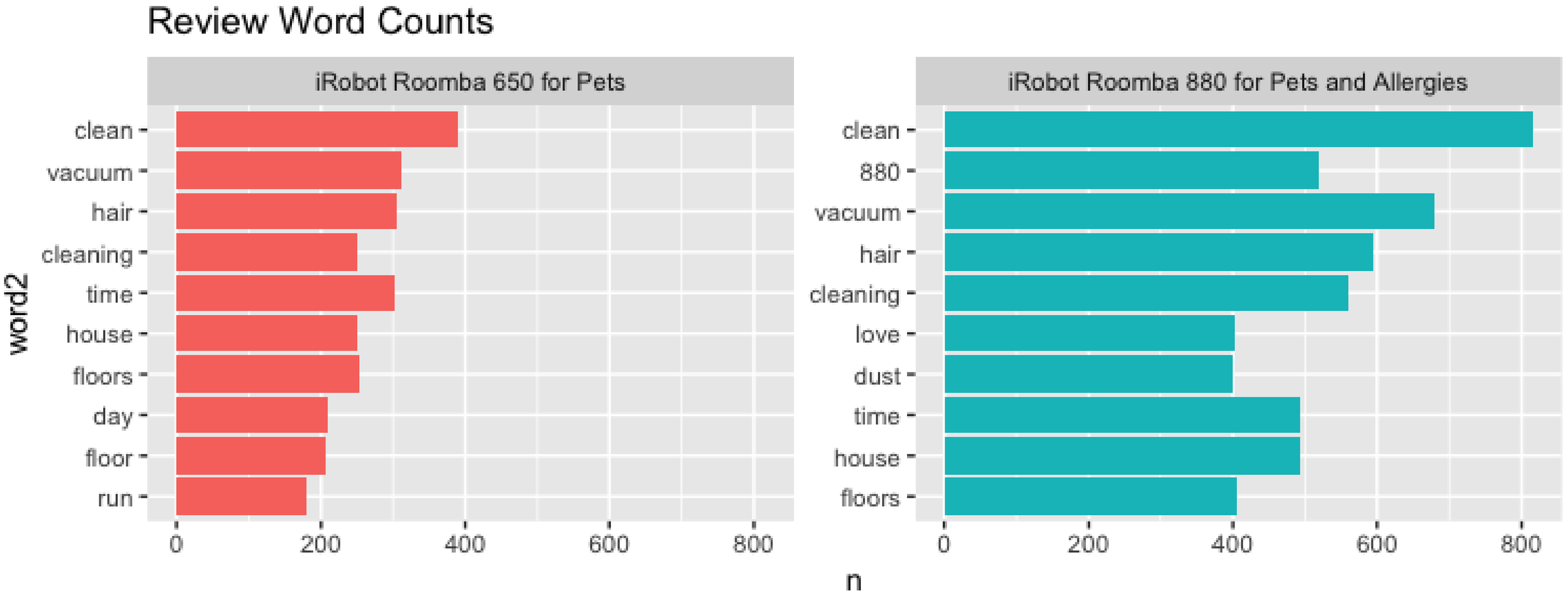
Using fct_reorder()

```
word_counts <- tidy_review %>%
  count(word, product) %>%
  group_by(product) %>%
  slice_max(n, n = 10) %>%
  ungroup() %>%
  mutate(word2 = fct_reorder(word, n))
```

Using facet_wrap()

```
ggplot(word_counts, aes(x = word2, y = n, fill = product)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ product, scales = "free_y") +
  coord_flip() +
  ggtitle("Review Word Counts")
```

Using facet_wrap()



Let's practice!

INTRODUCTION TO TEXT ANALYSIS IN R

Plotting word clouds

INTRODUCTION TO TEXT ANALYSIS IN R

Maham Faisal Khan
Senior Data Science Content Developer

Using wordcloud()

```
library(wordcloud)
word_counts <- tidy_review %>%
  count(word)
wordcloud(
  words = word_counts$word,
  freq = word_counts$n,
  max.words = 30
)
```

Fixed size and random start points

```
wordcloud(
  words = word_counts$word,
  freq = word_counts$n,
  max.words = 30
)
```

Number of words in the cloud

```
wordcloud(
  words = word_counts$word,
  freq = word_counts$n,
  max.words = 70
)
```

Using colors

```
wordcloud(
  words = word_counts$word,
  freq = word_counts$n,
  max.words = 30,
  colors = "blue"
)
```

Let's practice!

INTRODUCTION TO TEXT ANALYSIS IN R