FinalProject-TextMining

```r
library(tidytext)
library(tidyverse)
library(scales) # for previewing colors
```

1) Load in Data

```r
dryer <- read_csv("hair_dryer_verif.csv") # only interested in verified purchases
paci <- read_csv("pacifier_verif.csv")
micro <- read_csv("microwave_verif.csv")
head(dryer)
```

```
## # A tibble: 6 x 24
##   marketplace customer_id review_id product_id product_parent product_title
##   <chr>             <dbl> <chr>     <chr>               <dbl> <chr>
## 1 us             34678741 R9T1FE2Z~ B003V264WW      732252283 remington ac~
## 2 us             11599505 RE36JAD5~ B0009XH6V4      670161917 andis micro ~
## 3 us              2282190 RIDHM8B7~ B0007NZPY6       16483457 conair pro h~
## 4 us             43669858 R14QGWPC~ B00BB8ZIW0      253917972 remington si~
## 5 us             51995766 R230LCPQ~ B000065DJY      919751065 revlon 1875w~
## 6 us               180659 RY00YLVI~ B003FBG88E      195677102 conair pro s~
## # ... with 18 more variables: product_category <chr>, star_rating <dbl>,
## #   helpful_votes <dbl>, total_votes <dbl>, vine <chr>,
## #   verified_purchase <chr>, review_headline <chr>, review_body <chr>,
## #   review_date <date>, type <chr>, helpful_ratio <dbl>, helpful <lgl>,
## #   has_votes <lgl>, impact_pos <dbl>, impact_neg <dbl>, cum_star_rating <dbl>,
## #   review_month <date>, impact_star <chr>
```

```r
head(paci)
```

```
## # A tibble: 6 x 24
##   marketplace customer_id review_id product_id product_parent product_title
##   <chr>             <dbl> <chr>     <chr>               <dbl> <chr>
## 1 us             40626522 R1A3ZUBR~ B00793CZAE      572944212 mary meyer w~
## 2 us             16290022 RLJNYBK4~ B003PCYMP4      911821018 wubbanub lam~
## 3 us             10216509 R26QCW75~ B003CK3LDI      392768822 wubbanub inf~
## 4 us               114040 R2E7N0TV~ B003CK3LDI      392768822 wubbanub inf~
## 5 us             27971579 R1S09VMC~ B003PCYMP4      911821018 wubbanub lam~
## 6 us             36369192 R2EUVAGK~ B00PF841HA      449026476 philips aven~
## # ... with 18 more variables: product_category <chr>, star_rating <dbl>,
## #   helpful_votes <dbl>, total_votes <dbl>, vine <chr>,
## #   verified_purchase <chr>, review_headline <chr>, review_body <chr>,
## #   review_date <date>, type <chr>, helpful_ratio <dbl>, helpful <lgl>,
## #   has_votes <lgl>, impact_pos <dbl>, impact_neg <dbl>, cum_star_rating <dbl>,
## #   review_month <date>, impact_star <chr>
```

```
head(micro)
```

```
## # A tibble: 6 x 24
##   marketplace customer_id review_id product_id product_parent product_title
##   <chr>             <dbl> <chr>     <chr>               <dbl> <chr>
## 1 us             21879631 RY52KZAB~ B0052G14E8      423421857 danby 0.7 cu~
## 2 us             14964566 R3GCOEV4~ B0055UBB4O      423421857 danby 0.7 cu~
## 3 us             13230389 R1V2OPPN~ B0052G14E8      423421857 danby 0.7 cu~
## 4 us             43655888 R9Q0QDTL~ B004ZU09QQ      423421857 danby 0.7 cu~
## 5 us               117794 R3DL7HYC~ B005GSZB7I      827502283 whirlpool st~
## 6 us             16018452 R3M88678~ B004ZU09QQ      423421857 danby 0.7 cu~
## # ... with 18 more variables: product_category <chr>, star_rating <dbl>,
## #   helpful_votes <dbl>, total_votes <dbl>, vine <chr>,
## #   verified_purchase <chr>, review_headline <chr>, review_body <chr>,
## #   review_date <date>, type <chr>, helpful_ratio <dbl>, helpful <lgl>,
## #   has_votes <lgl>, impact_pos <dbl>, impact_neg <dbl>, cum_star_rating <dbl>,
## #   review_month <date>, impact_star <chr>
```

2) Data Preparation

Separate review headline, body, and product title into individual words to look for correlation between words and other factors

Define Stopwords

```
stop_words_es <- stopwords::stopwords("es", source = "stopwords-iso")
stop_words_fr <- stopwords::stopwords("fr", source = "stopwords-iso")
stop_words_en <- stopwords::stopwords("en", source = "stopwords-iso")
stop_words_extra <- c("i’m","we’re","it’s","you’re","that’s","they’ve")
stop_words_comp <- c(stop_words_es,stop_words_fr,stop_words_en,stop_words_extra)
stop_words_comp <- data.frame("word" = stop_words_comp, lexicon = 'SMART',stringsAsFactors = F)
```

function to unnest into indiv words

```
# df: dataframe that contains words
# words: words to unnest
# additional: additional columns to select
words_unnest <- function(df,words,additional){
  unnested <- df %>% unnest_tokens(word,!!words) %>% anti_join(stop_words_comp) %>% select(!!additional
  return(unnested)
}
```

Create base word dataframes to perform analysis on

```
# unnest captions into individual words
dryer_head <- words_unnest(dryer,quo(review_headline),quo(c(review_id,review_date,word,star_rating,help
dryer_body <- words_unnest(dryer,quo(review_body),quo(c(review_id,review_date,word,star_rating,helpful,
dryer_p_title <- words_unnest(dryer,quo(product_title),quo(c(review_id,review_date,word,star_rating,hel
# head(dryer_p_title)

micro_head <- words_unnest(micro,quo(review_headline),quo(c(review_id,review_date,word,star_rating,help
micro_body <- words_unnest(micro,quo(review_body),quo(c(review_id,review_date,word,star_rating,helpful,
micro_p_title <- words_unnest(micro,quo(product_title),quo(c(review_id,review_date,word,star_rating,hel

paci_head <- words_unnest(paci,quo(review_headline),quo(c(review_id,review_date,word,star_rating,helpful
paci_body <- words_unnest(paci,quo(review_body),quo(c(review_id,review_date,word,star_rating,helpful,vi
```

```r
paci_p_title <- words_unnest(paci,quo(product_title),quo(c(review_id,review_date,word,star_rating,helpfu
```

function to count words and find avg star rating per word

```r
# df: dataframe that contains words
words_count_rtg <- function(df){
  word_ct_rtg <- df %>% select(-c(review_id,review_date)) %>%
  group_by(word) %>% mutate(word_count=n(),avg_star = mean(star_rating)) %>% # find word count + avg st
  arrange(desc(word_count)) %>% select(-star_rating)

  word_ct_rtg <- word_ct_rtg[!(duplicated(word_ct_rtg$word)),] # remove duplicates of each word
  return(word_ct_rtg)
}
```

Find these values for header, body, and product title

```r
dryer_head_ct <- words_count_rtg(dryer_head)
dryer_body_ct <- words_count_rtg(dryer_body)
dryer_p_title_ct <- words_count_rtg(dryer_p_title)
```

Define frequent words as having counts > 5 appearances for review headers and product titles and > 200 appearances for review bodies (since we at least want to retain 1 quartile of data)

```r
dryer_head_ct <- dryer_head_ct %>% filter(word_count > 10) %>% arrange(desc(avg_star))
dryer_body_ct <- dryer_body_ct %>% filter(word_count > 200) %>% arrange(desc(avg_star))
dryer_p_title_ct <- dryer_p_title_ct %>% filter(word_count > 10) %>% arrange(desc(word_count))
dryer_head_ct$word <- factor(dryer_head_ct$word, levels = dryer_head_ct$word) # keeps order for barplot
dryer_body_ct$word <- factor(dryer_body_ct$word, levels = dryer_body_ct$word)
dryer_p_title_ct$word <- factor(dryer_p_title_ct$word, levels = dryer_p_title_ct$word)
head(dryer_body_ct)
```

```
## # A tibble: 6 x 5
## # Groups:   word [6]
##   word      helpful vine  word_count avg_star
##   <fct>     <lgl>   <chr>      <int>    <dbl>
## 1 loves     NA      n            222     4.86
## 2 highly    NA      n            287     4.80
## 3 love      NA      n           1898     4.79
## 4 excellent NA      n            306     4.75
## 5 perfect   NA      n            457     4.70
## 6 dries     NA      n            884     4.61
```

Repeat for Microwave

```r
micro_head_ct <- words_count_rtg(micro_head)
micro_body_ct <- words_count_rtg(micro_body)
micro_p_title_ct <- words_count_rtg(micro_p_title)
```

Not as many microwave reviews so lower cutoff for filter

```r
micro_head_ct <- micro_head_ct %>% filter(word_count > 2) %>% arrange(desc(avg_star))
micro_body_ct <- micro_body_ct %>% filter(word_count > 5) %>% arrange(desc(avg_star))
micro_p_title_ct <- micro_p_title_ct %>% filter(word_count > 2) %>% arrange(desc(word_count))
micro_head_ct$word <- factor(micro_head_ct$word, levels = micro_head_ct$word) # keeps order for barplot
micro_body_ct$word <- factor(micro_body_ct$word, levels = micro_body_ct$word)
micro_p_title_ct$word <- factor(micro_p_title_ct$word, levels = micro_p_title_ct$word)
```

Repeat for Pacifier

```r
paci_head_ct <- words_count_rtg(paci_head)
paci_body_ct <- words_count_rtg(paci_body)
paci_p_title_ct <- words_count_rtg(paci_p_title)
```

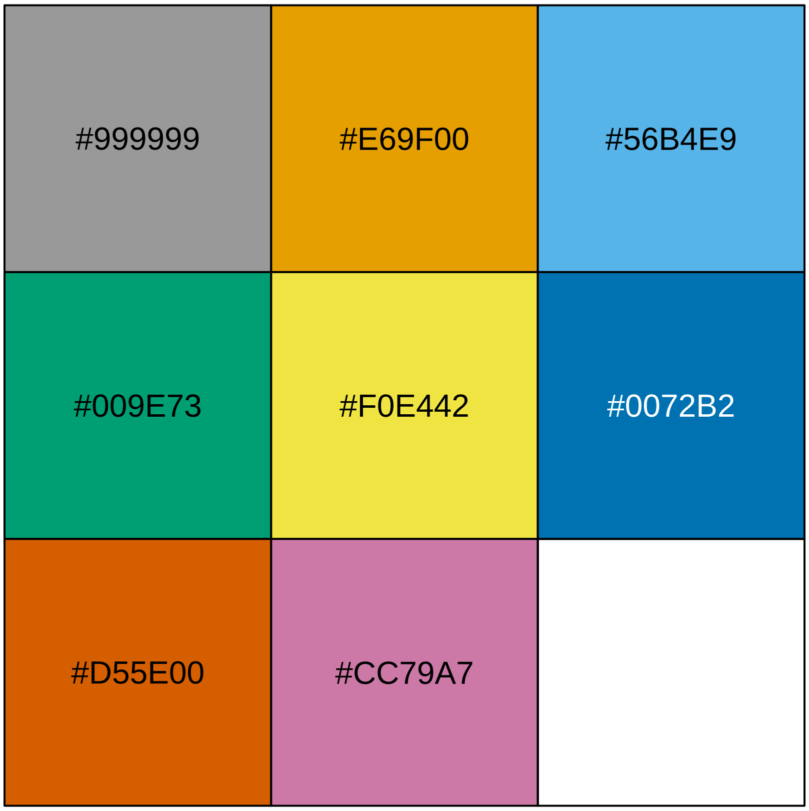
Lots of pacifier reviews so higher cutoff for filter

```r
paci_head_ct <- paci_head_ct %>% filter(word_count > 20) %>% arrange(desc(avg_star))
paci_body_ct <- paci_body_ct %>% filter(word_count > 200) %>% arrange(desc(avg_star))
paci_p_title_ct <- paci_p_title_ct %>% filter(word_count > 50) %>% arrange(desc(word_count))
paci_head_ct$word <- factor(paci_head_ct$word, levels = paci_head_ct$word) # keeps order for barplot
paci_body_ct$word <- factor(paci_body_ct$word, levels = paci_body_ct$word)
paci_p_title_ct$word <- factor(paci_p_title_ct$word, levels = paci_p_title_ct$word)
```

3) Graphs

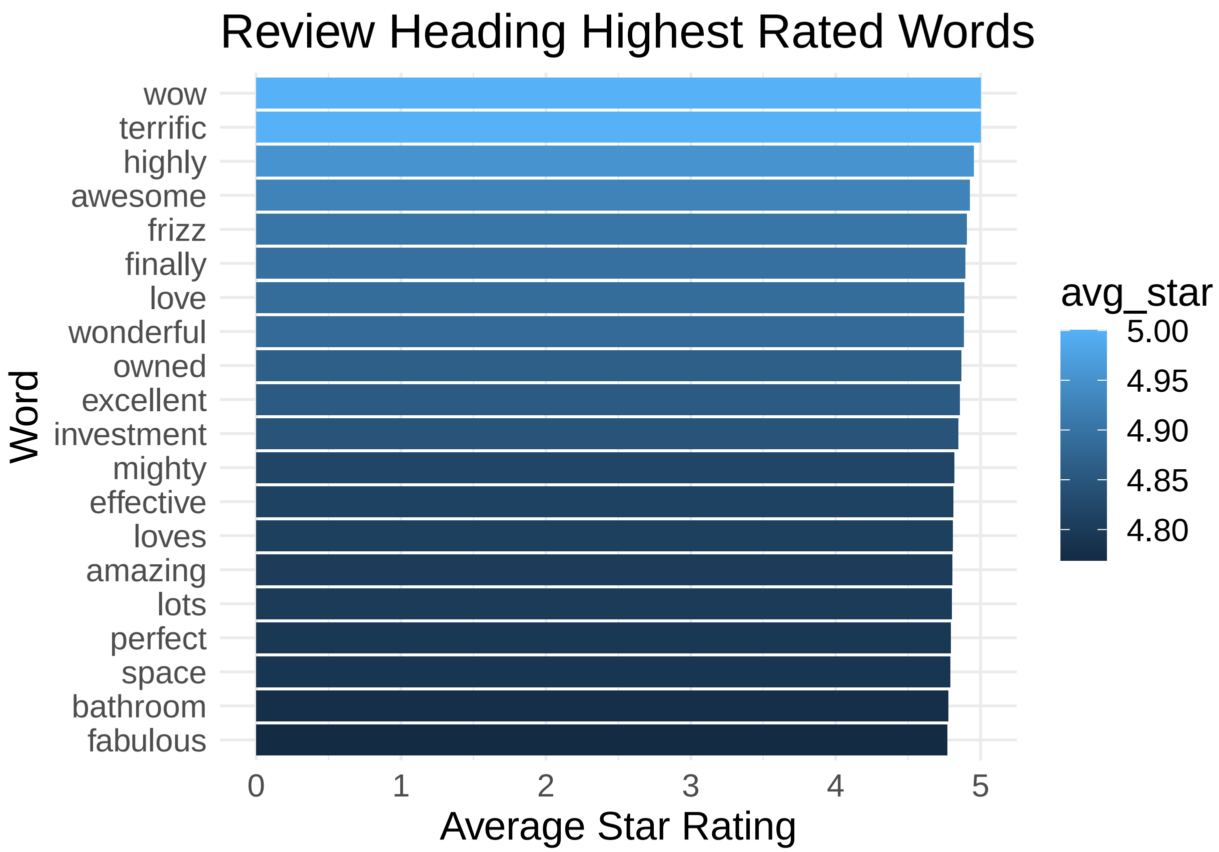
Define custom color sets

```r
cbp <- c("#999999", "#E69F00", "#56B4E9", "#009E73",
          "#F0E442", "#0072B2", "#D55E00", "#CC79A7")
show_col(cbp)
```



Still need to filter out words it seems…

```r
dh <- ggplot(head(dryer_head_ct,20),aes(x=reorder(word,avg_star),y=avg_star,fill=avg_star)) + geom_bar(
  theme_minimal() + theme(text = element_text(size=15)) + scale_fill_gradient() + coord_flip() +
  ggtitle("Review Heading Highest Rated Words") + ylab("Average Star Rating") + xlab("Word")
dh
```



Lots of words not related to features: wow, terrific, highly, wonderful, excellent, filter these out also in title, filter out nondescriptive words/pronouns -> make function

```
remove_nonsense <- function(df,nonsense){
  df <- df[!df$word %in% nonsense,]
  return(df)
}
```

```
nonsense <- c("wow","terrific","highly","awesome","finally","love","wonderful",
              "excellent","loves","amazing","lots","perfect","fabulous","favorite",
              "christmas","fantastic","super","glad","solid","satisfied","absolutely",
              "stars","happy","nice","recommend",
              "dissapointed","dissapointing","poor","horrible","star","amazon",
              "reviews","meh")
pronouns <- c("dryer","hair","conair","1875","andis","1875w","revlon","remington","1600","ac2015")
```

Refilter

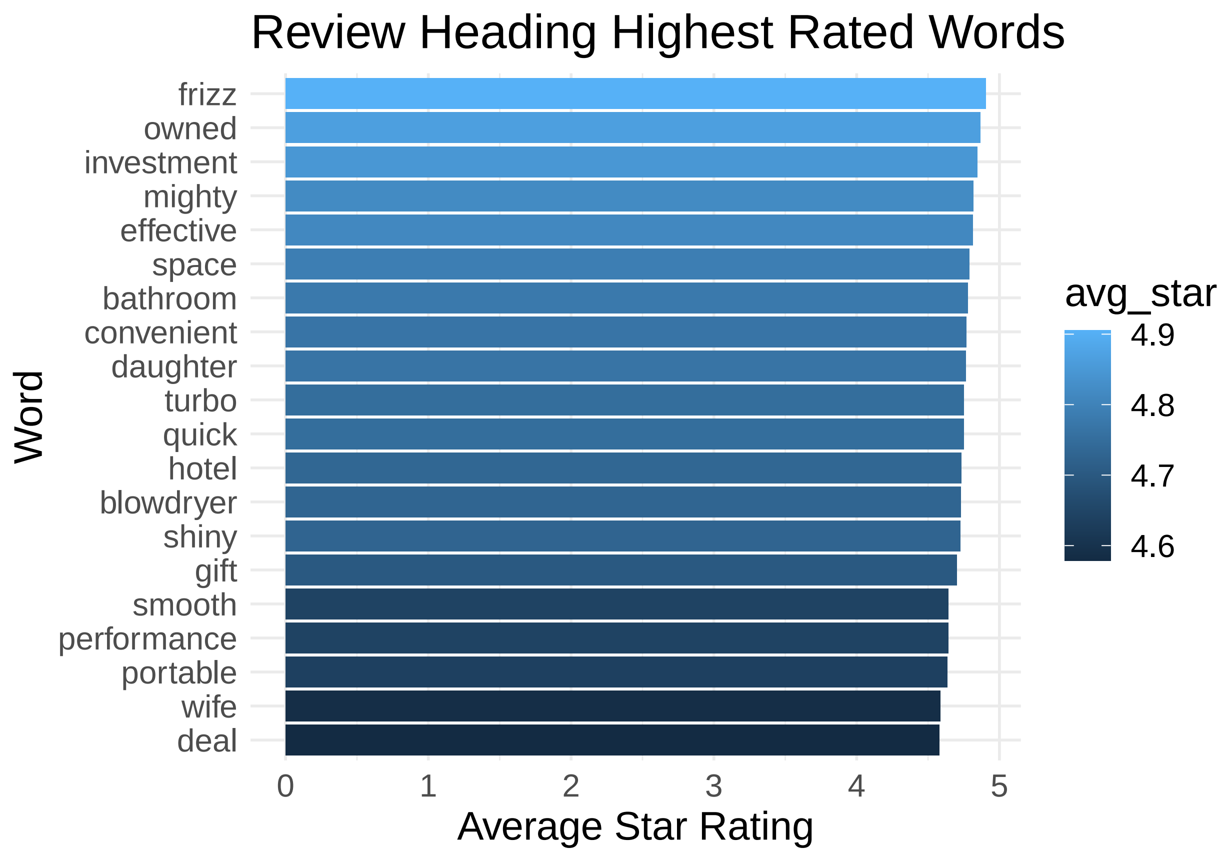
```
dryer_head_ct2 <- remove_nonsense(dryer_head_ct,nonsense)
dryer_body_ct2 <- remove_nonsense(dryer_body_ct,nonsense)
dryer_p_title_ct2 <- remove_nonsense(dryer_p_title_ct,nonsense)
```

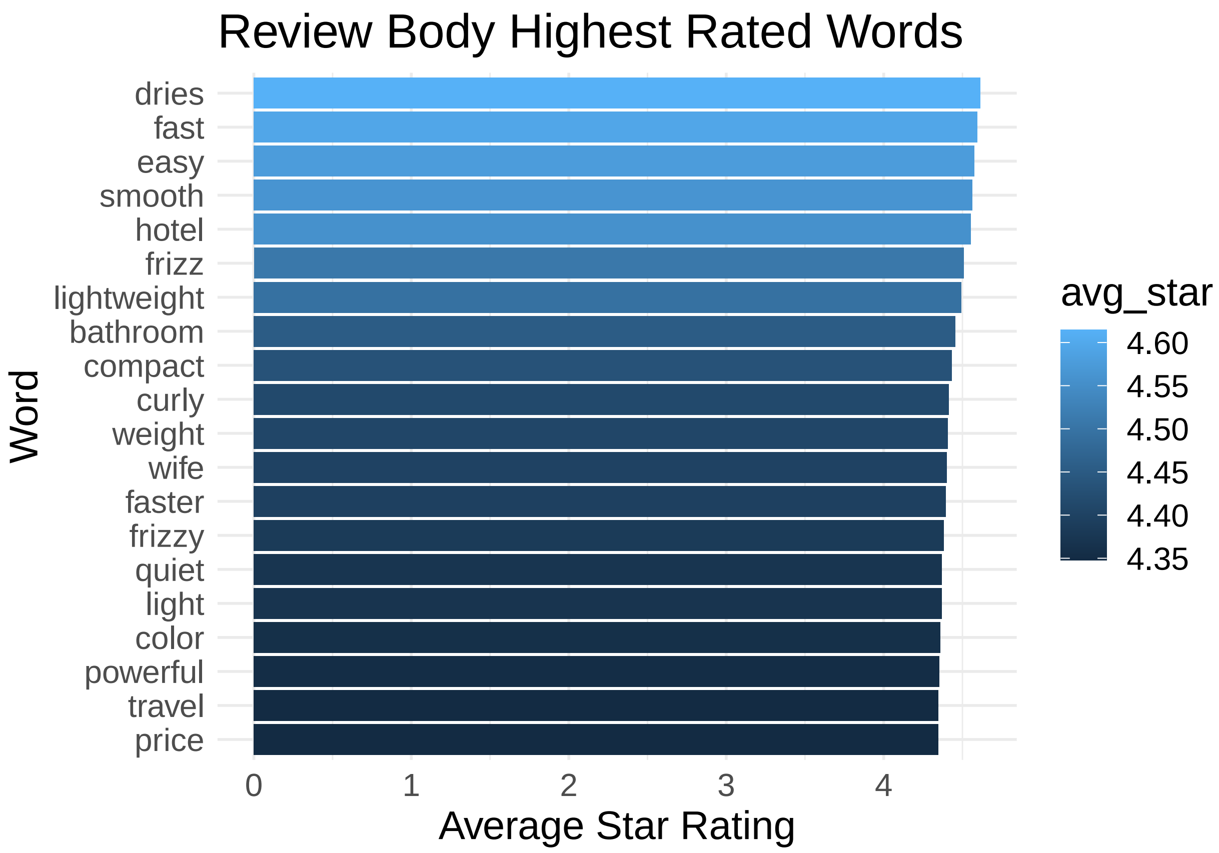
Try plotting again

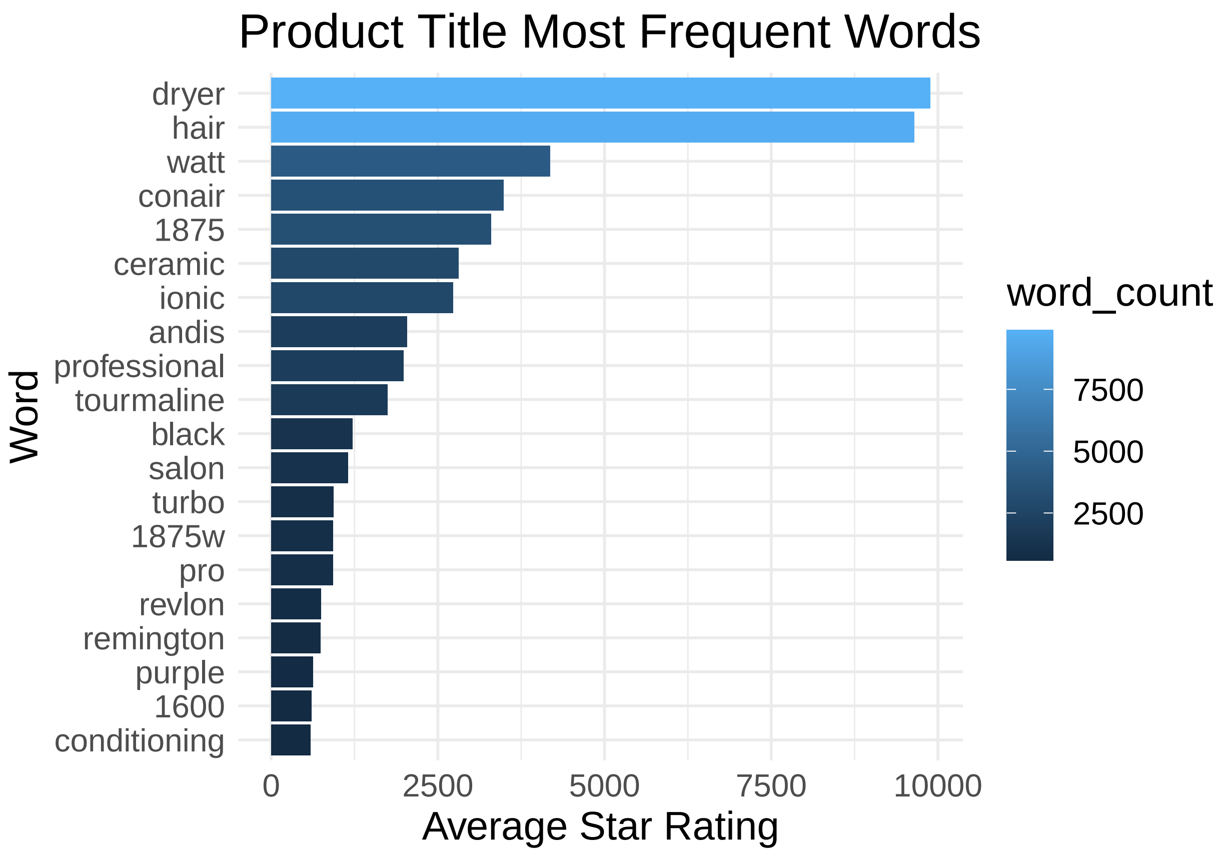
```
dh2 <- ggplot(head(dryer_head_ct2,20),aes(x=reorder(word,avg_star),y=avg_star,fill=avg_star)) + geom_ba
  theme_minimal() + theme(text = element_text(size=15)) + scale_fill_gradient() + coord_flip() +
  ggtitle("Review Heading Highest Rated Words") + ylab("Average Star Rating") + xlab("Word")
```

```
db2 <- ggplot(head(dryer_body_ct2,20),aes(x=reorder(word,avg_star),y=avg_star,fill=avg_star)) + geom_ba
  theme_minimal() + theme(text = element_text(size=15)) + scale_fill_gradient() + coord_flip() +
  ggtitle("Review Body Highest Rated Words") + ylab("Average Star Rating") + xlab("Word")

dpt2 <- ggplot(head(dryer_p_title_ct2,20),aes(x=reorder(word,word_count),y=word_count,fill=word_count))
  geom_bar(stat="identity") +
  theme_minimal() + theme(text = element_text(size=15)) + scale_fill_gradient() + coord_flip() +
  ggtitle("Product Title Most Frequent Words")+ ylab("Average Star Rating") + xlab("Word")
dh2; db2; dpt2
```



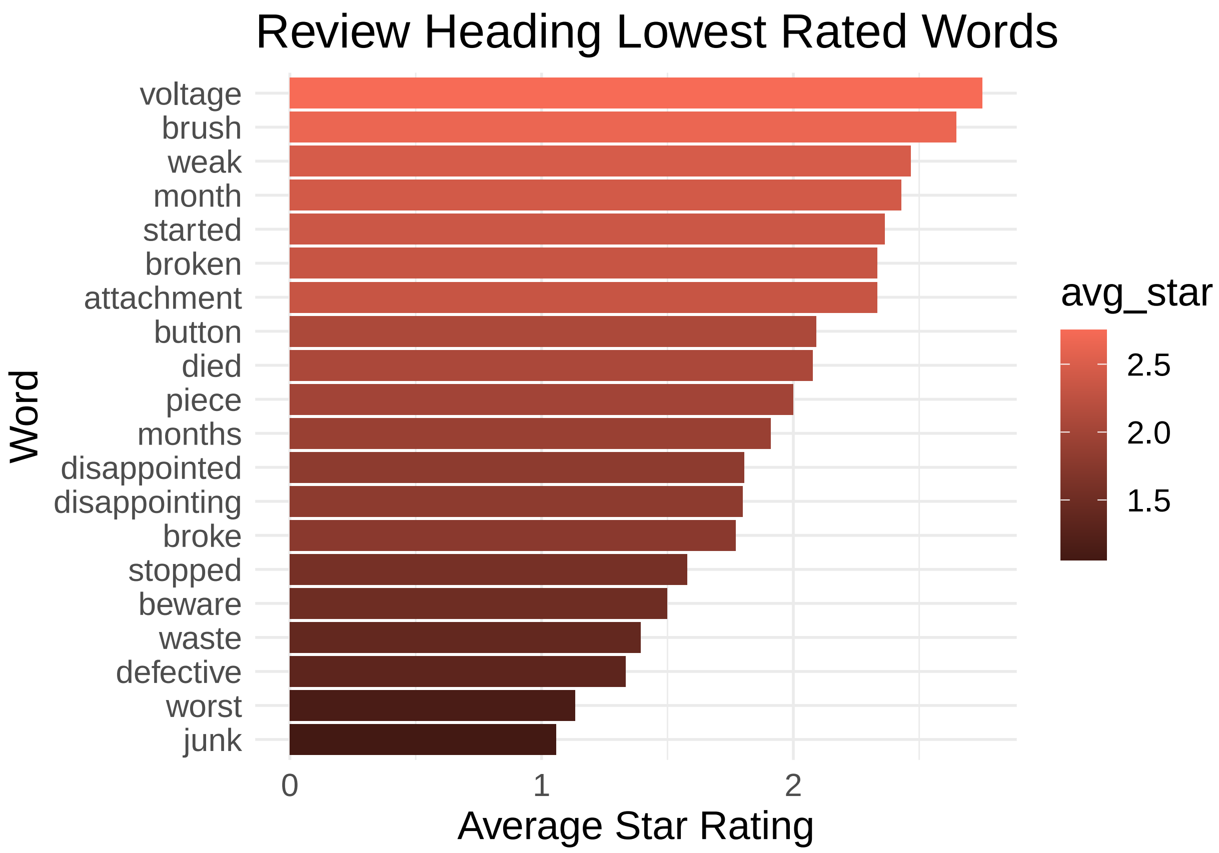

Review Body Highest Rated Words
Word
dries
fast
easy
smooth
hotel
frizz
lightweight
bathroom
compact
curly
weight
wife
faster
frizzy
quiet
light
color
powerful
travel
price
0
1
2
3
4
Average Star Rating
avg_star
4.60
4.55
4.50
4.45
4.40
4.35

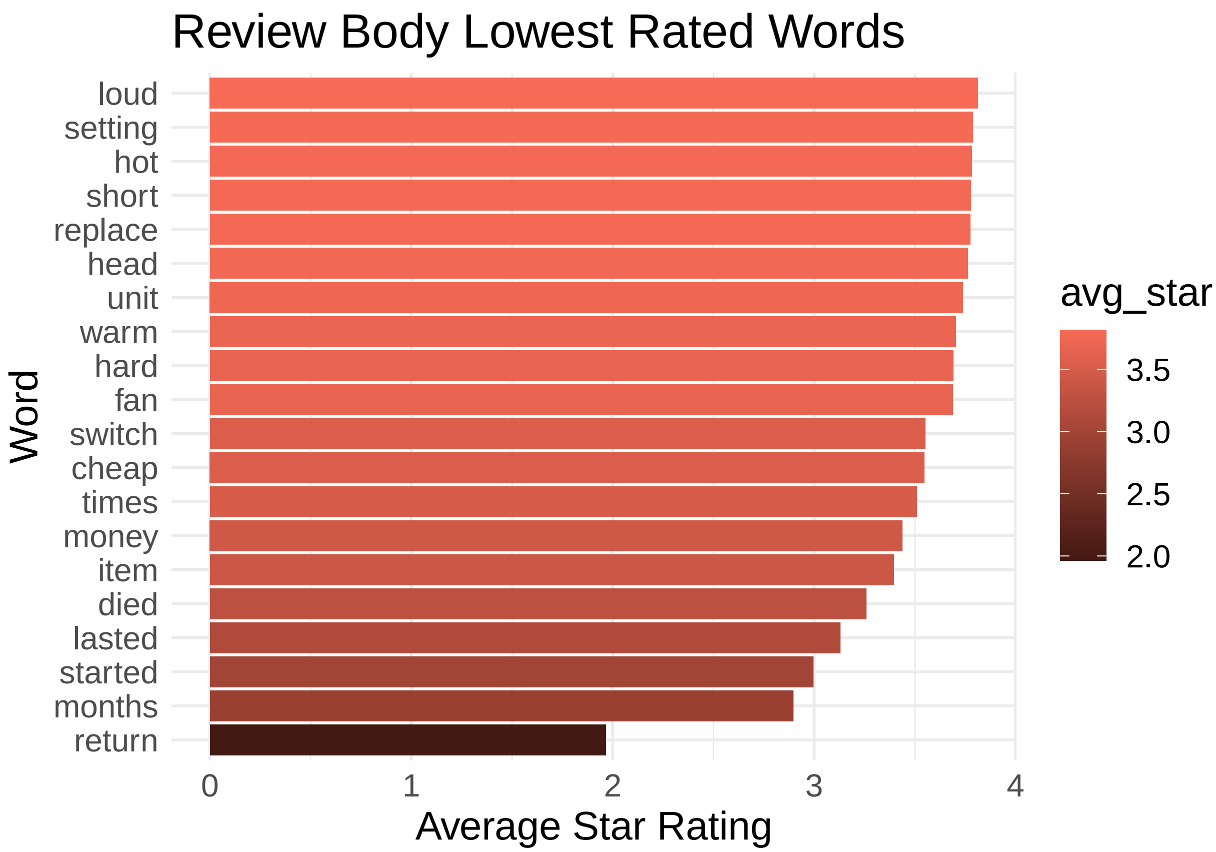


Look at lowest rated words

```
dh3 <- ggplot(tail(dryer_head_ct2,20),aes(x=reorder(word,avg_star),y=avg_star,fill=avg_star)) + geom_ba
  theme_minimal() + theme(text = element_text(size=15)) + scale_fill_gradient(low = "#431913",high = "#
  ggtitle("Review Heading Lowest Rated Words") + ylab("Average Star Rating") + xlab("Word")

db3 <- ggplot(tail(dryer_body_ct2,20),aes(x=reorder(word,avg_star),y=avg_star,fill=avg_star)) + geom_ba
  theme_minimal() + theme(text = element_text(size=15)) + scale_fill_gradient(low = "#431913",high = "#
  ggtitle("Review Body Lowest Rated Words") + ylab("Average Star Rating") + xlab("Word")
dh3; db3
```


Review Heading Lowest Rated Words
voltage
brush
weak
month
started
broken
attachment
button
died
piece
months
disappointed
disappointing
broke
stopped
beware
waste
defective
worst
junk
Word
0
1
2
Average Star Rating
avg_star
2.5
2.0
1.5



Repeat for Microwaves

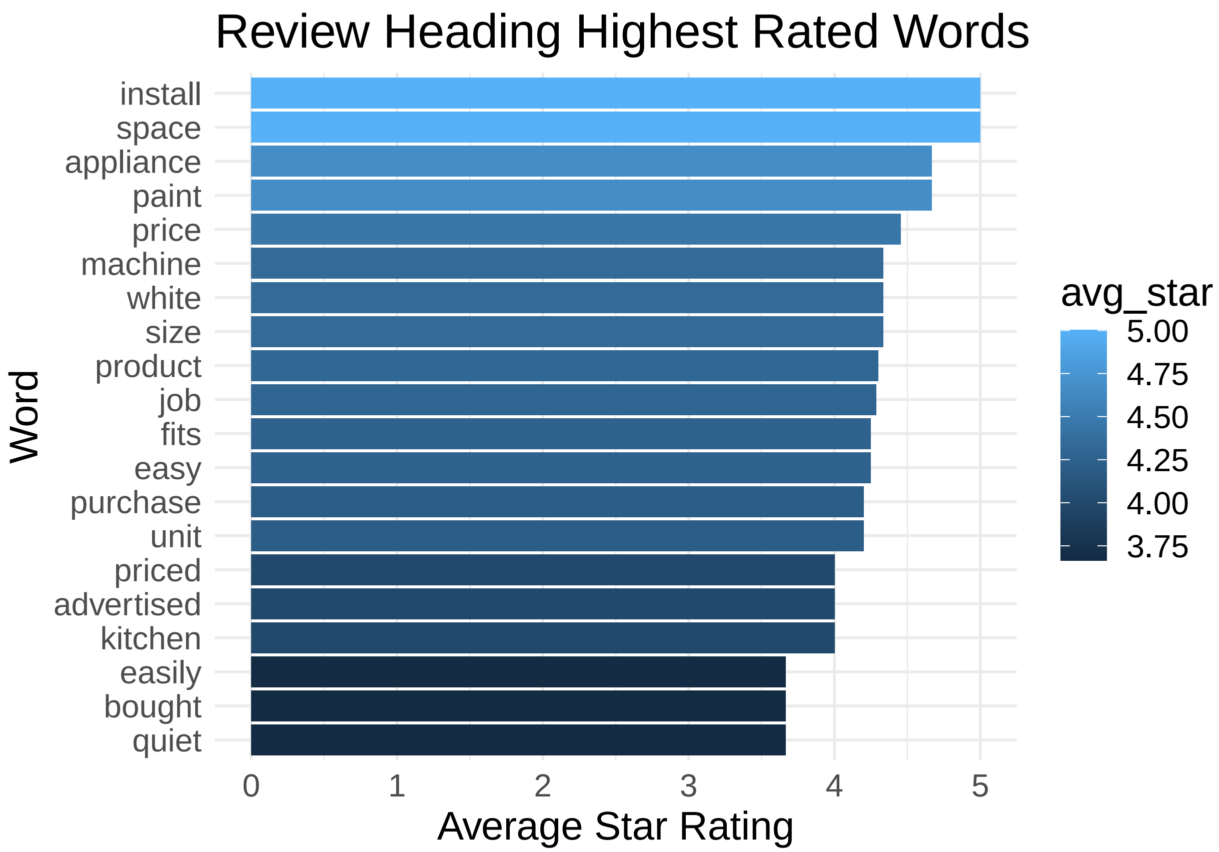
```
# add more nonsense words
nonsense <- c(nonsense,"expected","700","perfectly","limited","pleased","complaints","microwave","fine"
              "disappointed","recommended")
pronouns <- c(pronouns,"microwave","microwaves","countertop","danby","oven","range","98qbp0302","kit","
              "wmc20005yd","om75p","wb27x10017","wmc20005yw","mc11h6033ct","whirlpool","lg")
```

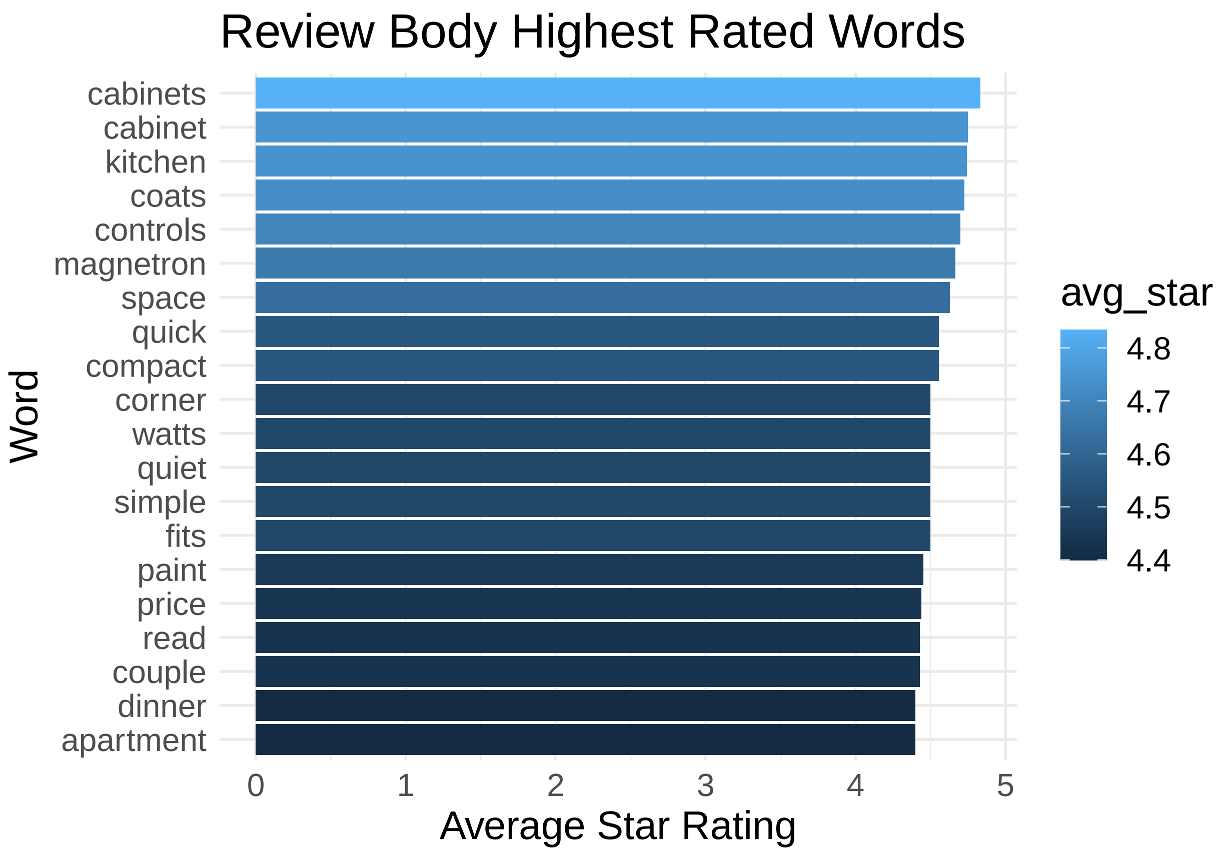
```
micro_head_ct2 <- remove_nonsense(micro_head_ct,nonsense)
micro_body_ct2 <- remove_nonsense(micro_body_ct,nonsense)
micro_p_title_ct2 <- remove_nonsense(micro_p_title_ct,pronouns)
```

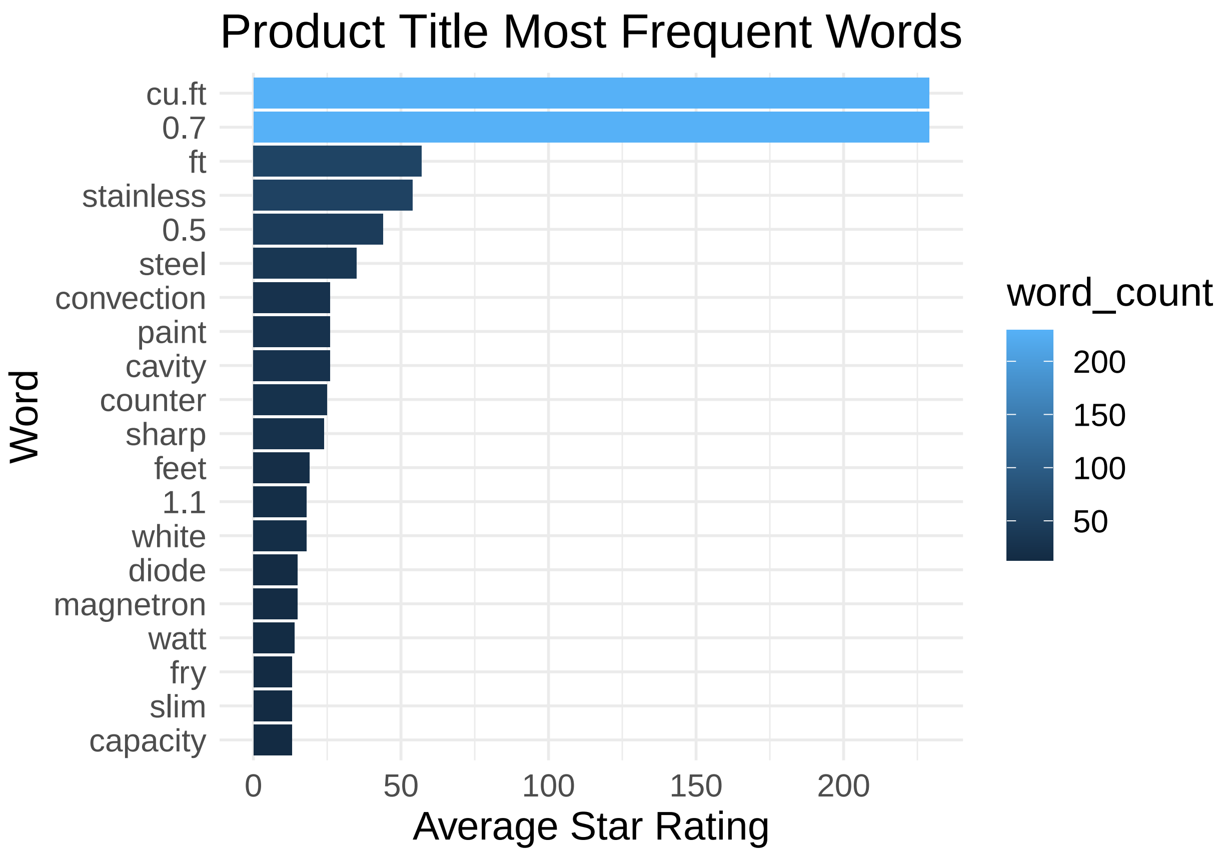
```
mh2 <- ggplot(head(micro_head_ct2,20),aes(x=reorder(word,avg_star),y=avg_star,fill=avg_star)) + geom_ba
  theme_minimal() + theme(text = element_text(size=15)) + scale_fill_gradient() + coord_flip() +
  ggtitle("Review Heading Highest Rated Words") + ylab("Average Star Rating") + xlab("Word")

mb2 <- ggplot(head(micro_body_ct2,20),aes(x=reorder(word,avg_star),y=avg_star,fill=avg_star)) + geom_ba
  theme_minimal() + theme(text = element_text(size=15)) + scale_fill_gradient() + coord_flip() +
  ggtitle("Review Body Highest Rated Words") + ylab("Average Star Rating") + xlab("Word")

mpt2 <- ggplot(head(micro_p_title_ct2,20),aes(x=reorder(word,word_count),y=word_count,fill=word_count))
  geom_bar(stat="identity") +
  theme_minimal() + theme(text = element_text(size=15)) + scale_fill_gradient() + coord_flip() +
  ggtitle("Product Title Most Frequent Words")+ ylab("Average Star Rating") + xlab("Word")
mh2; mb2; mpt2
```

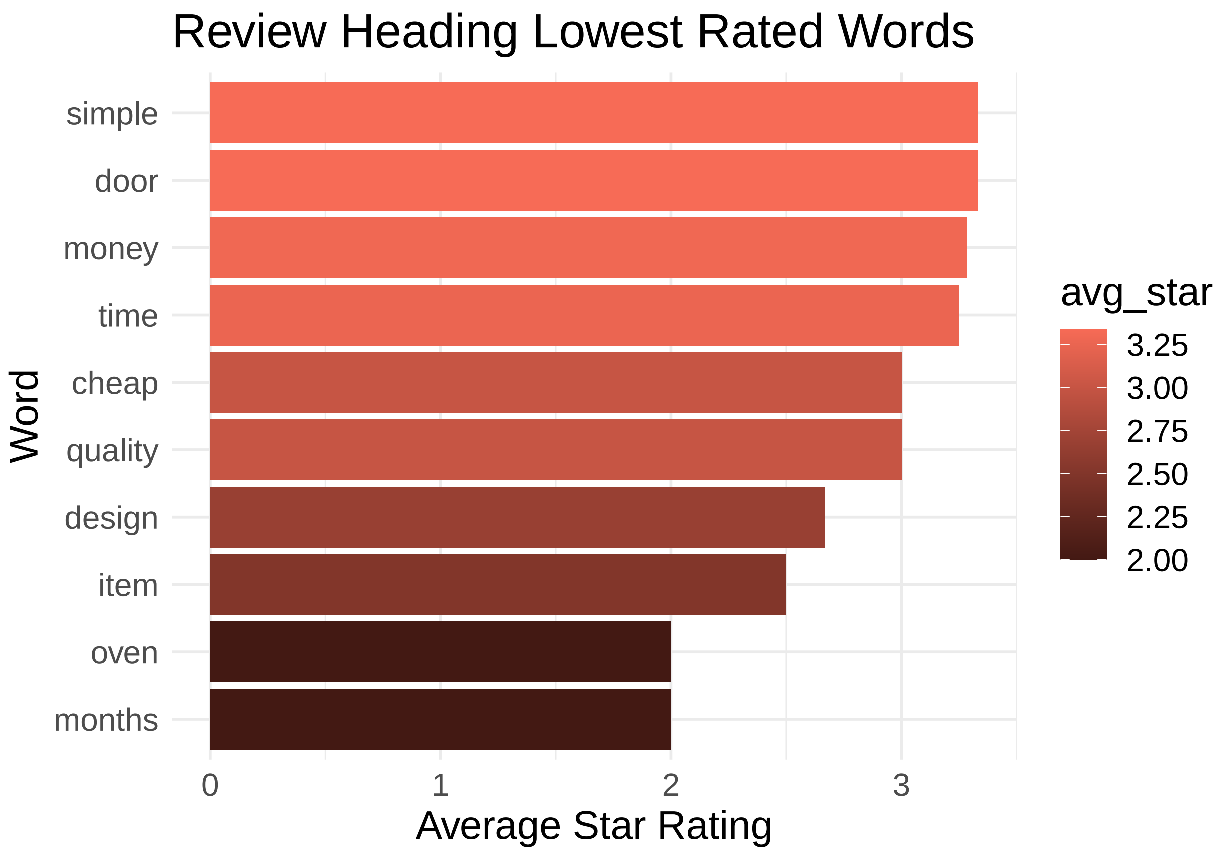

Review Heading Highest Rated Words
Word
install
space
appliance
paint
price
machine
white
size
product
job
fits
easy
purchase
unit
priced
advertised
kitchen
easily
bought
quiet
0
1
2
3
4
5
Average Star Rating
avg_star
5.00
4.75
4.50
4.25
4.00
3.75

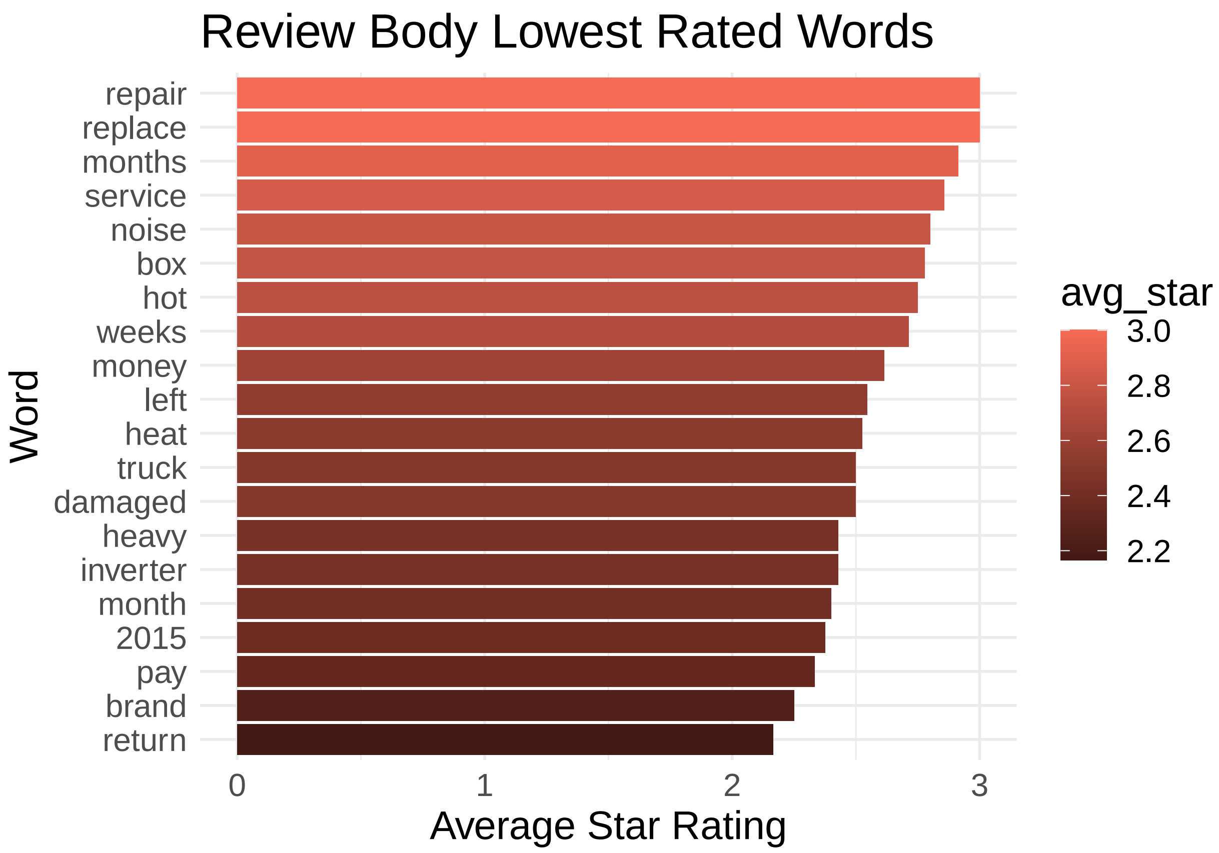

Review Body Highest Rated Words
cabinets
cabinet
kitchen
coats
controls
magnetron
space
quick
compact
corner
watts
quiet
simple
fits
paint
price
read
couple
dinner
apartment
Word
0
1
2
3
4
5
Average Star Rating
avg_star
4.8
4.7
4.6
4.5
4.4



```r
mh3 <- ggplot(tail(micro_head_ct2,10),aes(x=reorder(word,avg_star),y=avg_star,fill=avg_star)) + geom_ba
  theme_minimal() + theme(text = element_text(size=15)) + scale_fill_gradient(low = "#431913",high = "#
  ggtitle("Review Heading Lowest Rated Words") + ylab("Average Star Rating") + xlab("Word")

mb3 <- ggplot(tail(micro_body_ct2,20),aes(x=reorder(word,avg_star),y=avg_star,fill=avg_star)) + geom_ba
  theme_minimal() + theme(text = element_text(size=15)) + scale_fill_gradient(low = "#431913",high = "#
  ggtitle("Review Body Lowest Rated Words") + ylab("Average Star Rating") + xlab("Word")
mh3; mb3
```


Review Heading Lowest Rated Words
Word
simple
door
money
time
cheap
quality
design
item
oven
months
0
1
2
3
Average Star Rating
avg_star
3.25
3.00
2.75
2.50
2.25
2.00



Repeat for Pacifiers

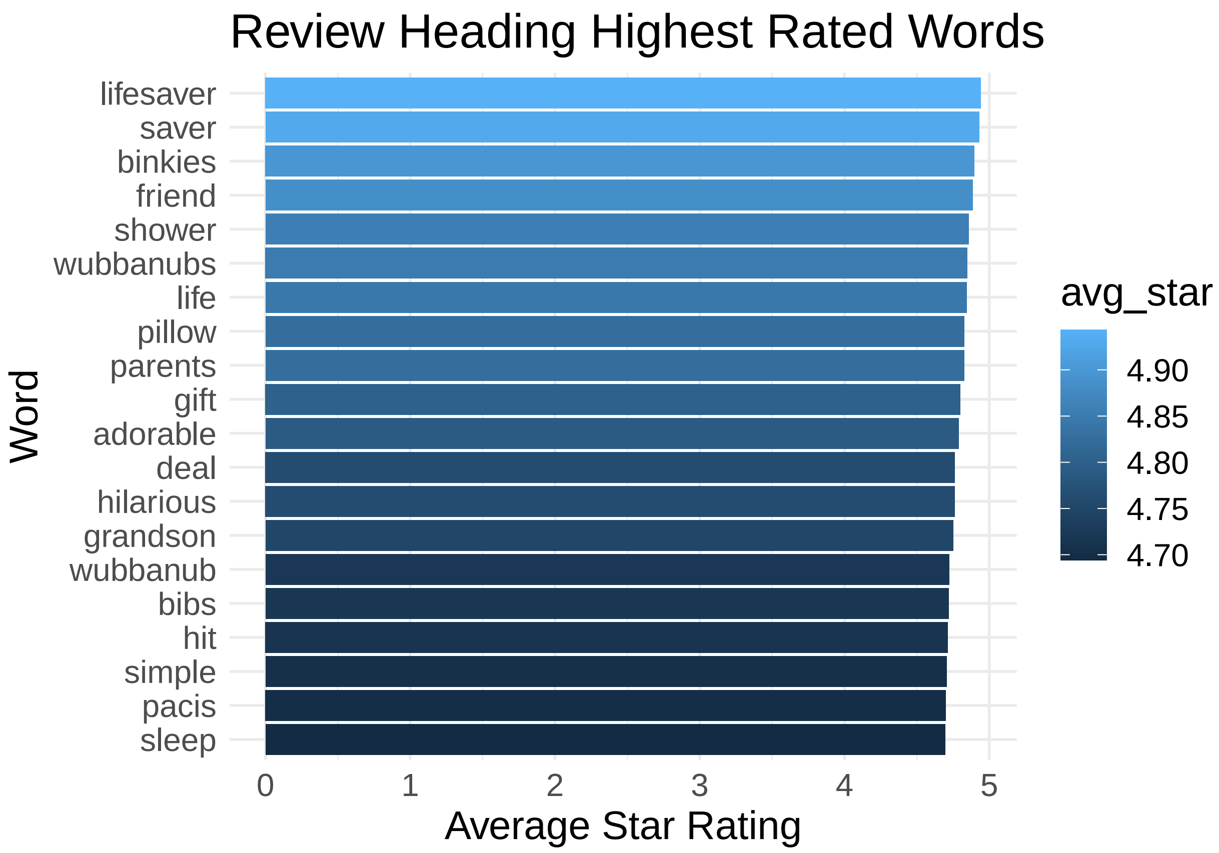
```
pronouns <- c(pronouns,"pacifier","pacifiers")
```

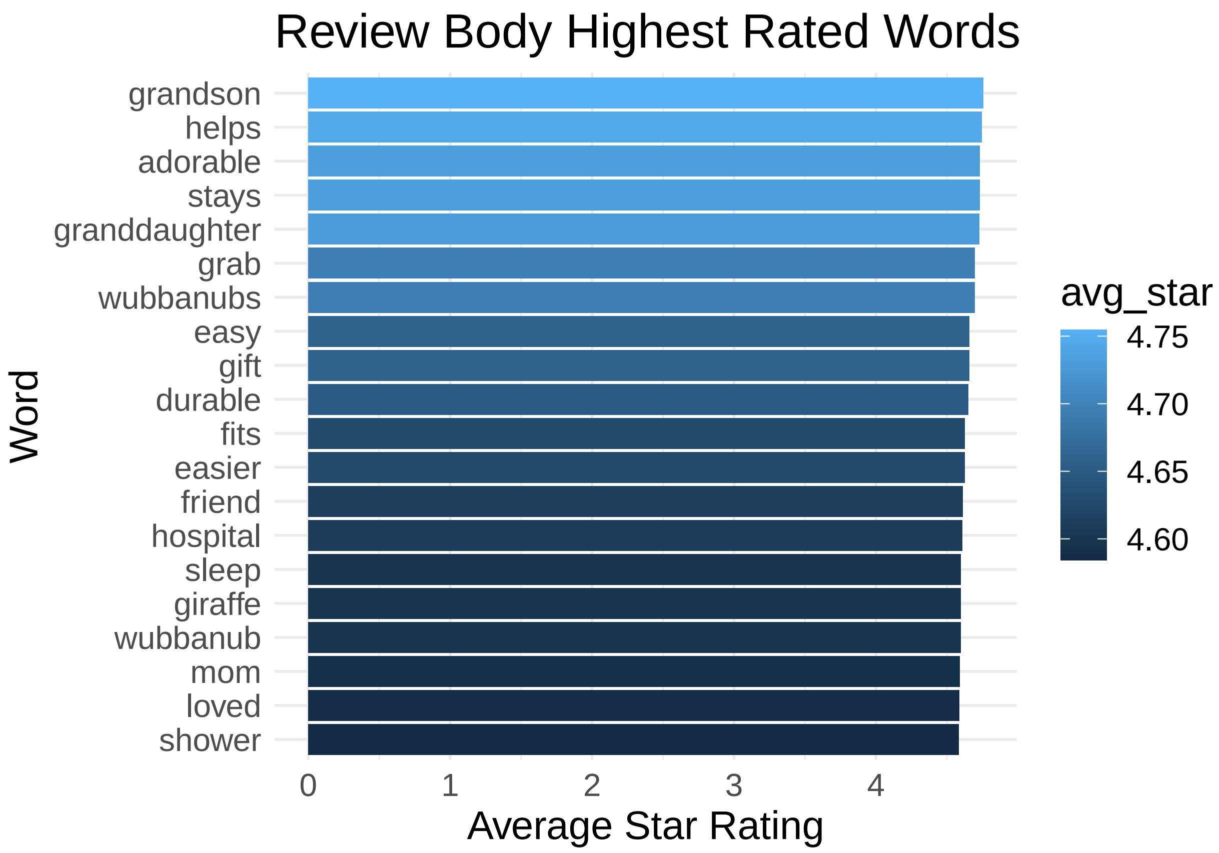
```
paci_head_ct2 <- remove_nonsense(paci_head_ct,nonsense)
paci_body_ct2 <- remove_nonsense(paci_body_ct,nonsense)
paci_p_title_ct2 <- remove_nonsense(paci_p_title_ct,pronouns)
```

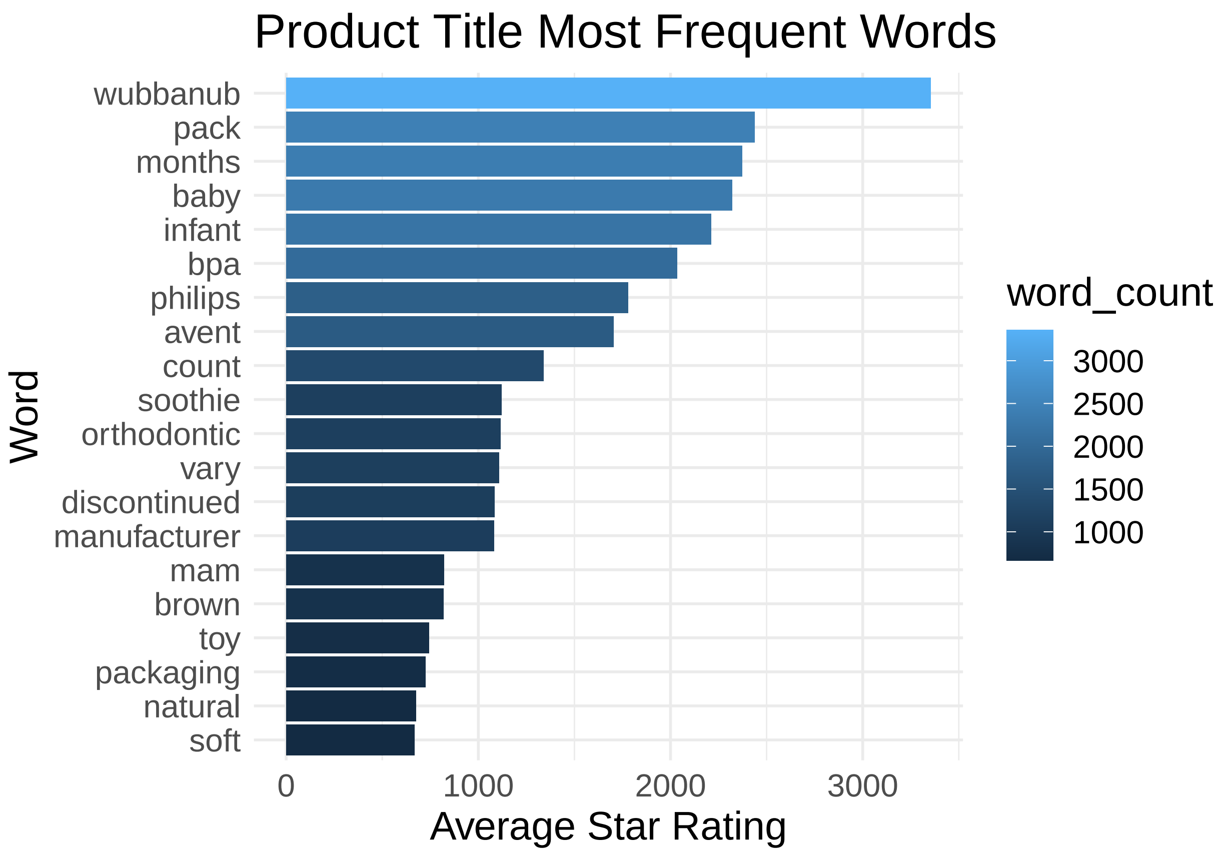
```
ph2 <- ggplot(head(paci_head_ct2,20),aes(x=reorder(word,avg_star),y=avg_star,fill=avg_star)) + geom_bar
  theme_minimal() + theme(text = element_text(size=15)) + scale_fill_gradient() + coord_flip() +
  ggtitle("Review Heading Highest Rated Words") + ylab("Average Star Rating") + xlab("Word")

pb2 <- ggplot(head(paci_body_ct2,20),aes(x=reorder(word,avg_star),y=avg_star,fill=avg_star)) + geom_bar
  theme_minimal() + theme(text = element_text(size=15)) + scale_fill_gradient() + coord_flip() +
  ggtitle("Review Body Highest Rated Words") + ylab("Average Star Rating") + xlab("Word")

ppt2 <- ggplot(head(paci_p_title_ct2,20),aes(x=reorder(word,word_count),y=word_count,fill=word_count))
  geom_bar(stat="identity") +
  theme_minimal() + theme(text = element_text(size=15)) + scale_fill_gradient() + coord_flip() +
  ggtitle("Product Title Most Frequent Words")+ ylab("Average Star Rating") + xlab("Word")
ph2; pb2; ppt2
```

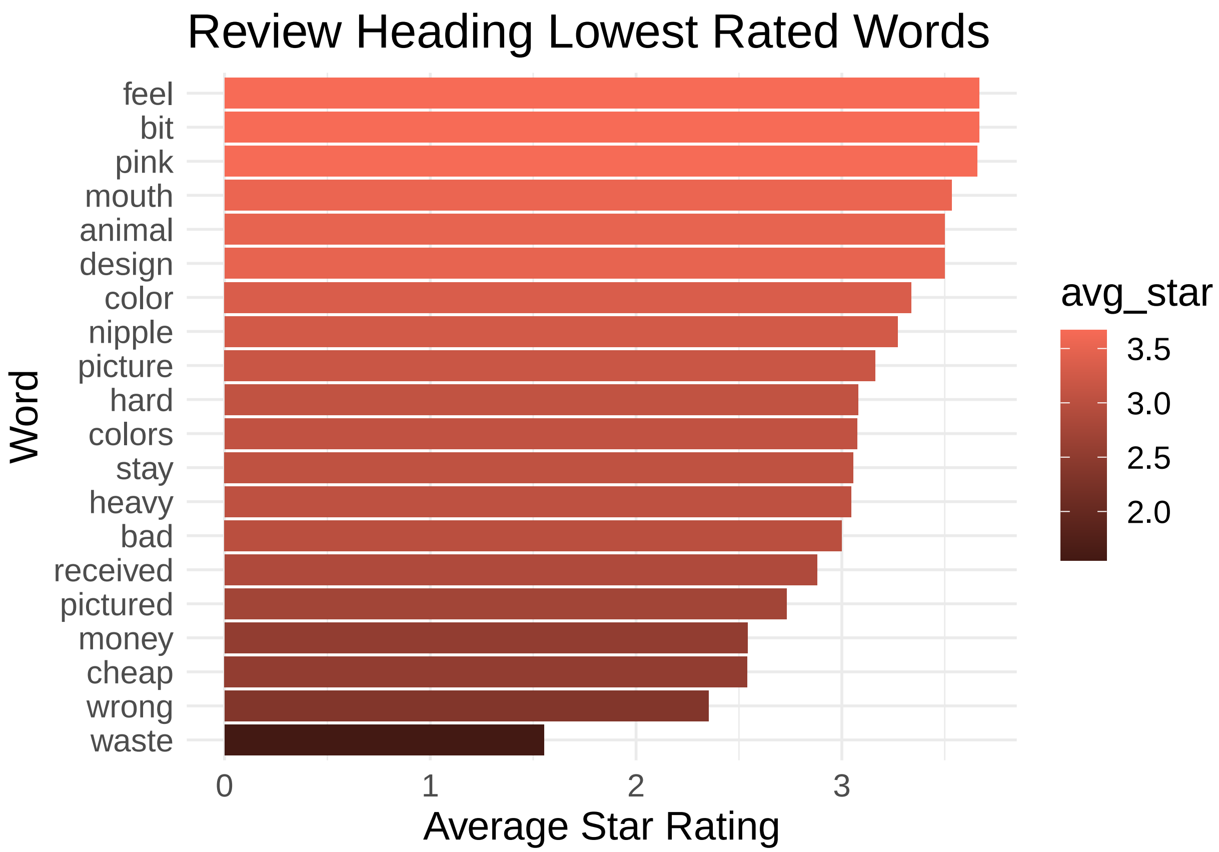

Review Heading Highest Rated Words
Word
lifesaver
saver
binkies
friend
shower
wubbanubs
life
pillow
parents
gift
adorable
deal
hilarious
grandson
wubbanub
bibs
hit
simple
pacis
sleep
0
1
2
3
4
5
Average Star Rating
avg_star
4.90
4.85
4.80
4.75
4.70

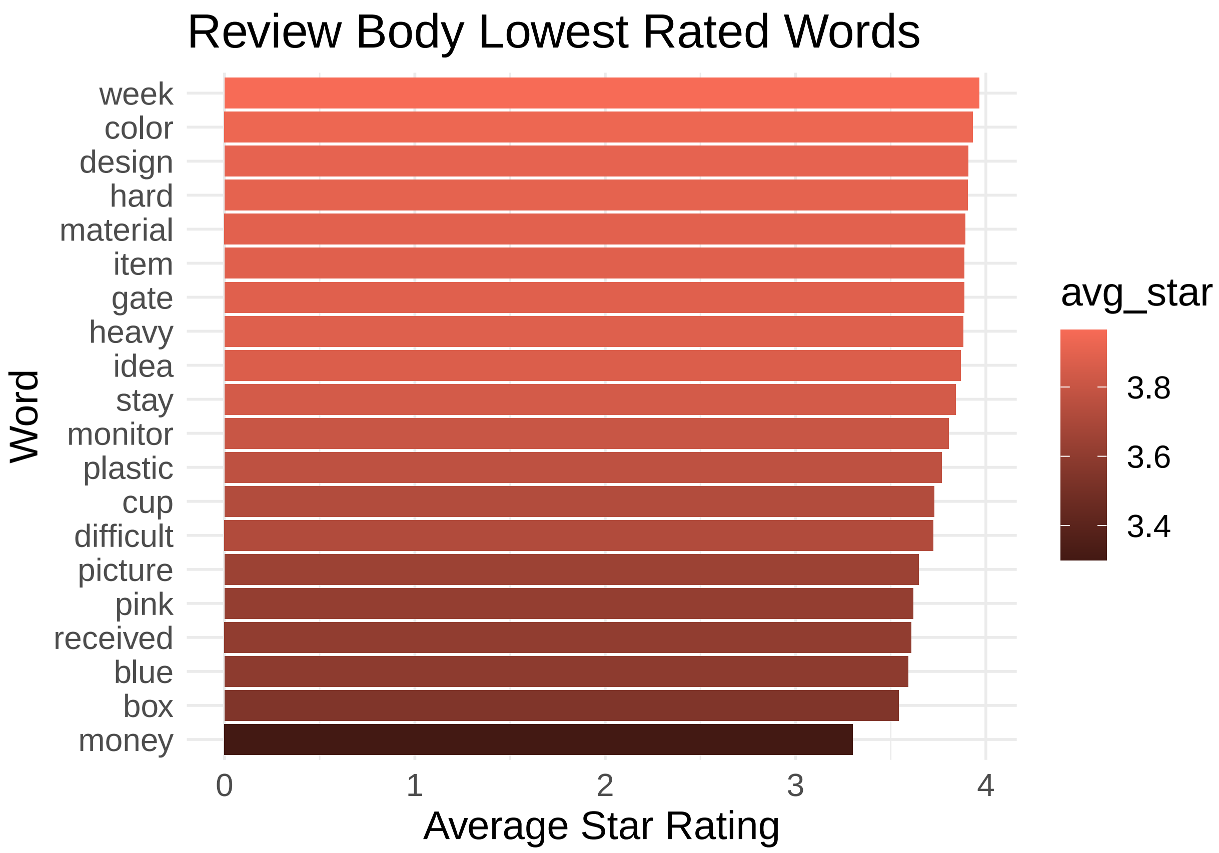

Review Body Highest Rated Words
Word
grandson
helps
adorable
stays
granddaughter
grab
wubbanubs
easy
gift
durable
fits
easier
friend
hospital
sleep
giraffe
wubbanub
mom
loved
shower
0
1
2
3
4
Average Star Rating
avg_star
4.75
4.70
4.65
4.60



```
ph3 <- ggplot(tail(paci_head_ct2,20),aes(x=reorder(word,avg_star),y=avg_star,fill=avg_star)) + geom_bar
  theme_minimal() + theme(text = element_text(size=15)) + scale_fill_gradient(low = "#431913",high = "#
  ggtitle("Review Heading Lowest Rated Words") + ylab("Average Star Rating") + xlab("Word")

pb3 <- ggplot(tail(paci_body_ct2,20),aes(x=reorder(word,avg_star),y=avg_star,fill=avg_star)) + geom_bar
  theme_minimal() + theme(text = element_text(size=15)) + scale_fill_gradient(low = "#431913",high = "#
  ggtitle("Review Body Lowest Rated Words") + ylab("Average Star Rating") + xlab("Word")
ph3; pb3
```


Review Heading Lowest Rated Words
Word
feel
bit
pink
mouth
animal
design
color
nipple
picture
hard
colors
stay
heavy
bad
received
pictured
money
cheap
wrong
waste
0
1
2
3
Average Star Rating
avg_star
3.5
3.0
2.5
2.0


Review Body Lowest Rated Words
Word
week
color
design
hard
material
item
gate
heavy
idea
stay
monitor
plastic
cup
difficult
picture
pink
received
blue
box
money
0
1
2
3
4
Average Star Rating
avg_star
3.8
3.6
3.4