Project 1

1. Introduction

Happiness is an emotion that we all as humans have the ability to experience, yet there are a plethora of things that uniquely bring each of us joy. The HappyDB database is a compendium of 100,000 different responses when describing a moment or experience that made them feel happy. The data is divided into seven categories that are quite general. A deeper dive into some of these categories can help highlight the commonality or differences in what brings each of us happiness!

2. Dataset

```
happy <- read.csv("/Users/daniellesolomon/Documents/GitHub/ads-spring2024-project1-ds3714/output/proces
head(happy)
```

```
##    hmid  wid reflection_period
## 1 27673 2053               24h
## 2 27674    2               24h
## 3 27675 1936               24h
## 4 27676  206               24h
## 5 27677 6227               24h
## 6 27678   45               24h
##
## 1                                                    I went on a successful date with someone I fe
## 2                                                                 I was happy when my son
## 3                                                                      I went to th
## 4 We had a serious talk with some friends of ours who have been flaky lately. They understood and we
## 5                                                         I went with grandchildren to butterfly
## 6
##
## 1                                                    I went on a successful date with someone I fe
## 2                                                                 I was happy when my son
## 3                                                                      I went to th
## 4 We had a serious talk with some friends of ours who have been flaky lately. They understood and we
## 5                                                         I went with grandchildren to butterfly
## 6
##   modified num_sentence ground_truth_category predicted_category id
## 1     TRUE            1                  <NA>          affection  1
## 2     TRUE            1                  <NA>          affection  2
## 3     TRUE            1                  <NA>           exercise  3
## 4     TRUE            2              bonding            bonding  4
## 5     TRUE            1                  <NA>          affection  5
## 6     TRUE            1              leisure            leisure  6
##                                                                  text
## 1 happy ago yesterday lot today months month happier happiest last week past
## 2 happy ago yesterday lot today months month happier happiest last week past
## 3 happy ago yesterday lot today months month happier happiest last week past
```

```
## 4 happy ago yesterday lot today months month happier happiest last week past
## 5 happy ago yesterday lot today months month happier happiest last week past
## 6 happy ago yesterday lot today months month happier happiest last week past
```

```
str(happy) # info on the characteristics of the data type in each column
```

```
## 'data.frame':    100392 obs. of  11 variables:
##  $ hmid                 : int  27673 27674 27675 27676 27677 27678 27679 27680 27681 27682 ...
##  $ wid                  : int  2053 2 1936 206 6227 45 195 740 3 4833 ...
##  $ reflection_period    : chr  "24h" "24h" "24h" "24h" ...
##  $ original_hm          : chr  "I went on a successful date with someone I felt sympathy and connecti
##  $ cleaned_hm           : chr  "I went on a successful date with someone I felt sympathy and connecti
##  $ modified             : logi  TRUE TRUE TRUE TRUE TRUE TRUE ...
##  $ num_sentence         : int  1 1 1 2 1 1 1 1 1 1 ...
##  $ ground_truth_category: chr  NA NA NA "bonding" ...
##  $ predicted_category   : chr  "affection" "affection" "exercise" "bonding" ...
##  $ id                   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ text                 : chr  "happy ago yesterday lot today months month happier happiest last weel
```

This data set was produced from the text processing R file that was provided in the class repository

This data set has 100,392 rows and 11 columns. The main columns of relevance for this project are "cleaned_hm" which contains the response sentences and "predicted_categories" which breaks the categorizes the responses into seven different groups based on relevant characteristics (words) that are present in the response sentence.

Note: When trying to reproduce these results by running the R code on one's device, the path file needs to be changed to wherever it is downloaded on the individual's device.

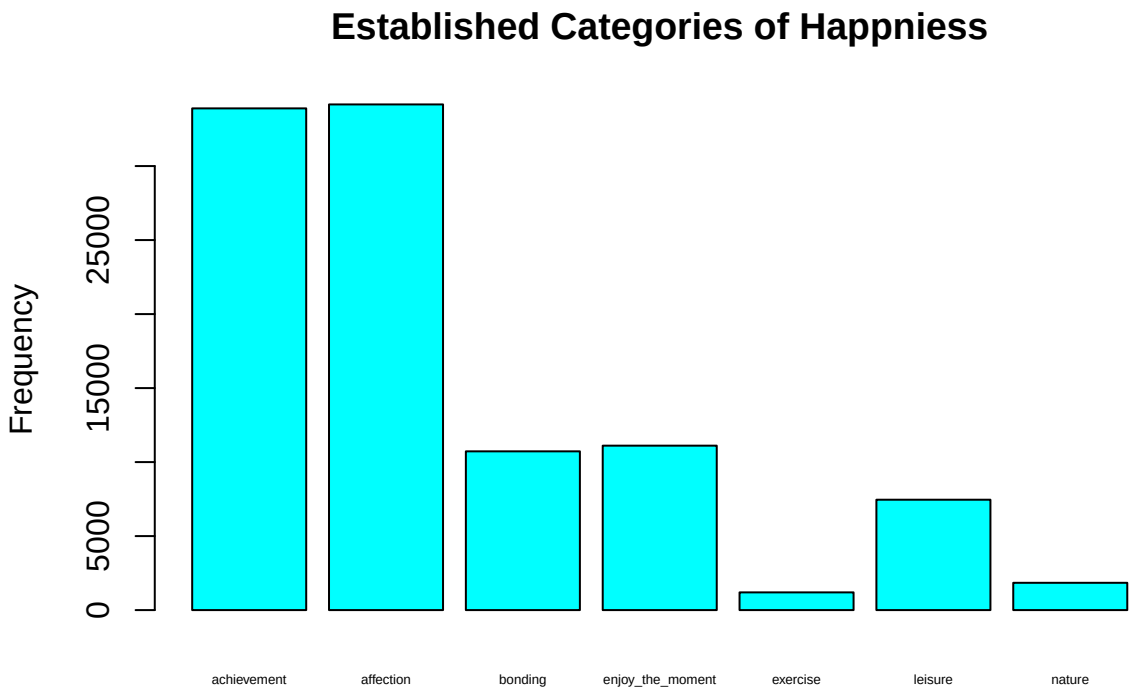
```
library(stringr)
library(stringi)
library(tm)
library(tidytext)
library(tidyverse)
library(DT)
library(ggplot2)
```

3. Explore!

Given categories

In the dataset, there are seven established categories that all of these responese are categorized into: achievement, affection, bonding, enjoy the moment, exercise, leisure, and nature.

```
t1 <- table(happy$predicted_category)
barplot(t1,
        xlab = "Category of Happiness",
        ylab = "Frequency",
        main = "Established Categories of Happniess",
        cex.names = 0.4,
        col = "cyan")
```



The category of affection seems to take the lead on most common category that brings people happiness, though it is closely followed by achievement.

Different categories of things that can make one happy

Although the data is broken into specific categories already, it may be interesting to explore different, more specific, categories based on how often certain words appear in this collection of responses. The given categories are quite broad and can even have overlap in some cases, so this exploration can be viewed as a deeper look into some of these categories, to see what they are really composed of. For simplicity's sake, the most common forms of the words of these categories are used in this search.

Preparation: Split strings so that each word is an element

```
sentences <- happy$cleaned_hm
words <- str_split_fixed(sentences, pattern = " ", n = Inf) # breaks up sentences into words - separate
length(words)
```

```
## [1] 115952760
```

```
str(words)
```

```
##  chr [1:100392, 1:1155] "I" "I" "I" "We" "I" "I" "I" "I" "YESTERDAY" ...
```

```
words <- as.character(stri_remove_empty(words, na_empty = FALSE))
length(words)
```

```
## [1] 1837370
```

There are 115952760 words in this entire dataset!

The words from the entire dataset will be considered as a whole in this further exploration of responses to see the existence of these topics outside of the already delineated categories.

Self/Posession (Achievement)

Happiness, as aforementioned, can be brought upon by a plethora of different things. The category of achievement does not clarify what this sense of achievement is brought upon by. This achievement can come from oneself or it can come from watching something or someone else achieve something.

The word "I" is a simple indicator that insinuates that the happiness is something that an individual feels. The word "my" is a possessive word that can insinuate that this happiness can come from someothing or someone belonging (to some degree) to that individual.

```
my <- str_detect(words, "my+|My+") # this function detects the occurences of the item in quotes
sum(my)
```

```
## [1] 73510
```

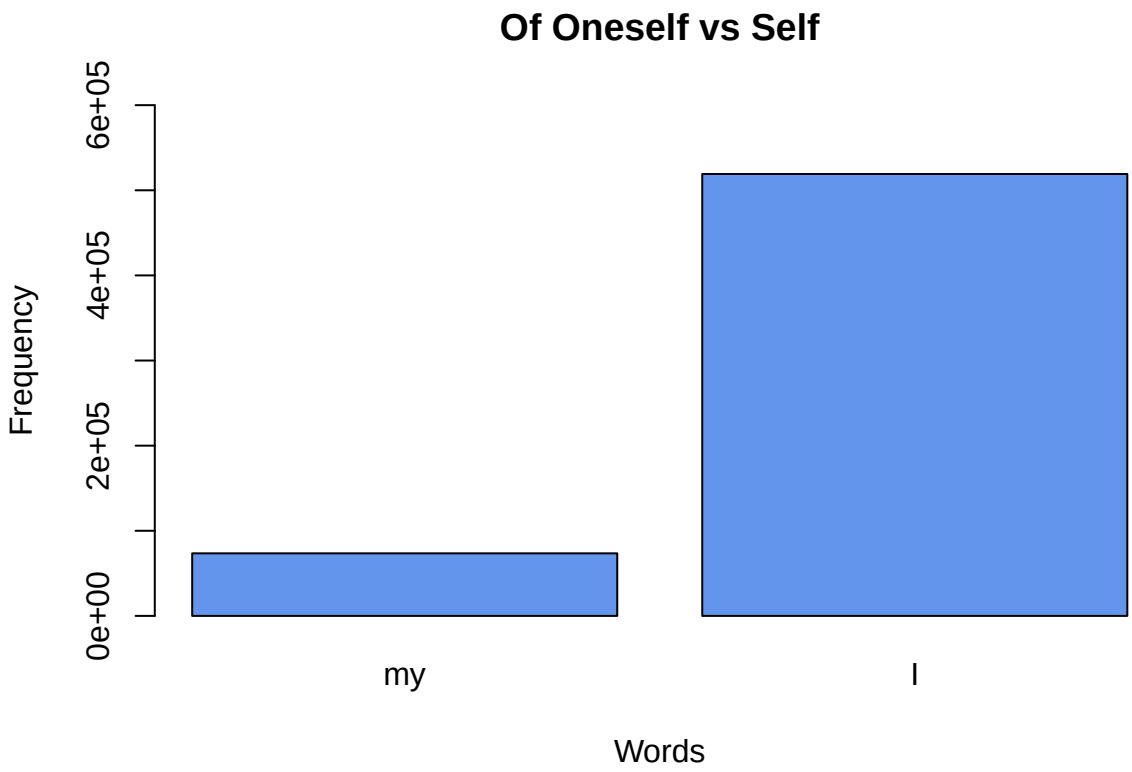
```
I <- str_detect(words, "I|i")
sum(I)
```

```
## [1] 519135
```

```
self <-data.frame(
  words = c("my", "I"),
  occurences = c(sum(my), sum(I)))    # this creates a dataframe of the occurrences of the relevant wor
str(self)
```

```
## 'data.frame':    2 obs. of  2 variables:
##  $ words     : chr  "my" "I"
##  $ occurences: int  73510 519135
```

```
barplot(height = self$occurence, name = self$words, ylim=c(0,600000),xlab = "Words", ylab = "Frequency"
```



In general, it seems that overwhelmingly more people find happineses something that they did themselves rather than observing something from someone/something that belongs to them!

Family/Pets (Affection)

From the original categories, we see that affection is the most popular category of response when asked about a source of happiness. Affection is quite a broad term. We can take a look at what this category could possibly be composed of.

Family and pets are two very popular examples of possible sources of affection. I broke down these categories into the most common associated terms to see how they appear in this dataset.

```
# family
brother <- str_detect(words, "brother")
sum(brother)
```

```
## [1] 1494
```

```
sister <- str_detect(words, "sister")
sum(sister)
```

```
## [1] 1728
```

```
sibling <- str_detect(words, "sibling+|Sibling+")
sum(sibling)
```

```
## [1] 88
```

```
mom <- str_detect(words, "mom|Mom|mother|Mother")
sum(mom)
```

```
## [1] 6797
```

```
dad <- str_detect(words, "dad|Dad|Father|father")
sum(dad)
```

```
## [1] 1943
```

```
parents <- str_detect(words, "parent+|Parent+")
sum(parents)
```

```
## [1] 1341
```

```
son <- str_detect(words, "son+|Son+")
sum(son)
```

```
## [1] 6784
```

```
daughter <- str_detect(words, "daughter+|Daughter+")
sum(daughter)
```

```
## [1] 3526
```

```
child <- str_detect(words, "kid+|Kid+|child|Child|baby|Baby|children|Children")
sum(child)
```

```
## [1] 4438
```

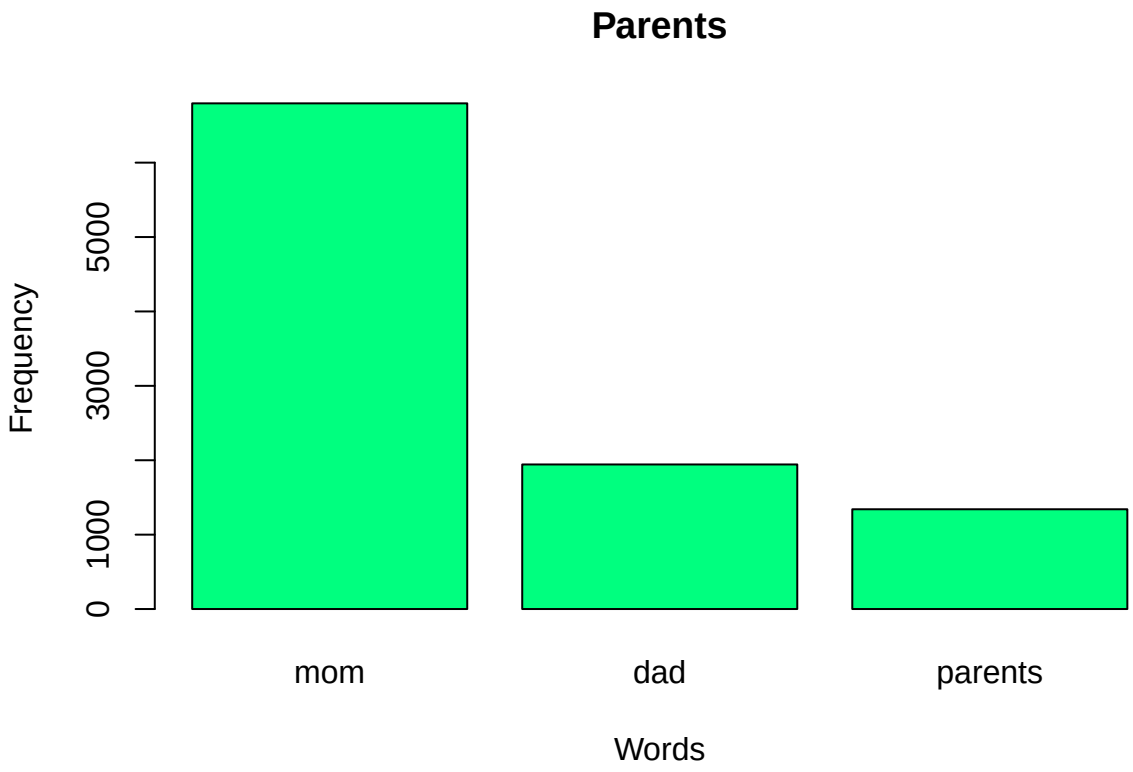
```
spouse <- str_detect(words, "wife|husband|Wife|Husband|Spouse|spouse")
sum(spouse)
```

```
## [1] 5719
```

```
partner <- str_detect(words, "girlfriend|Girlfriend|boyfriend|Boyfriend|partner|Partner")
sum(partner)
```

```
## [1] 3642
```

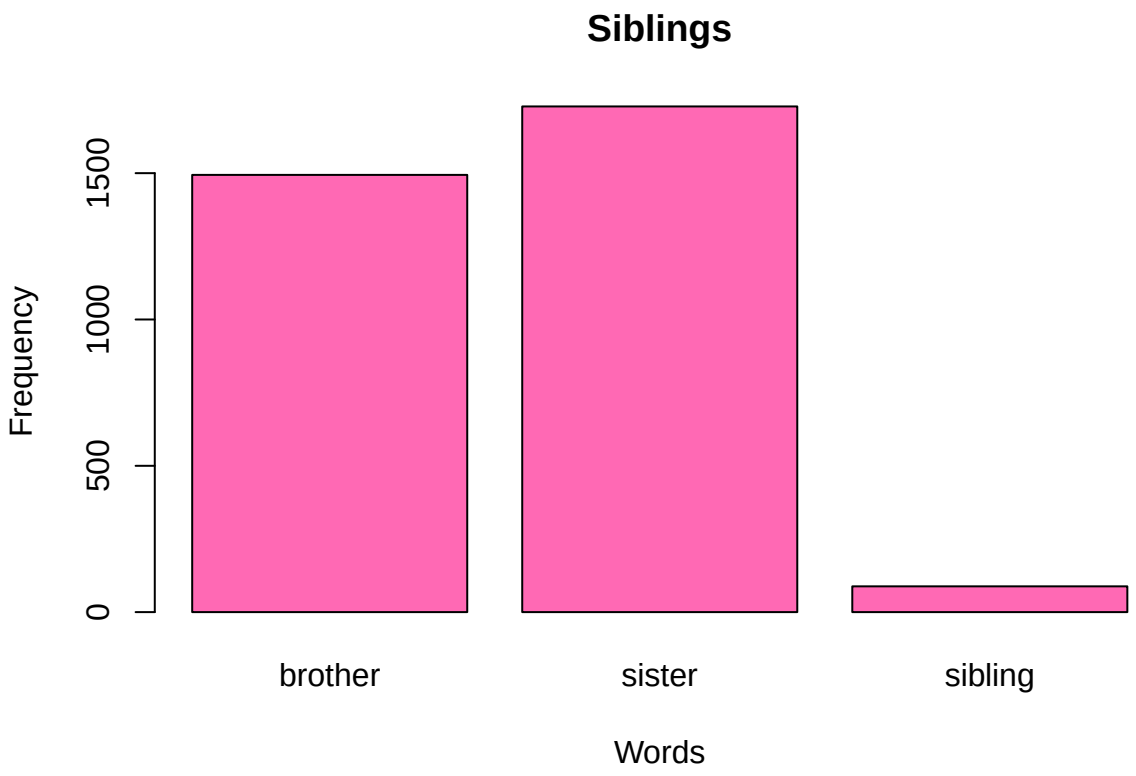
```r
parent <- data.frame(
  words = c("mom", "dad", "parents"),
  occurence = c(sum(mom), sum(dad), sum(parents)))
barplot(height = parent$occurence, name = parent$words,
        xlab = "Words", ylab = "Frequency", main = "Parents",
        col = "springgreen")
```



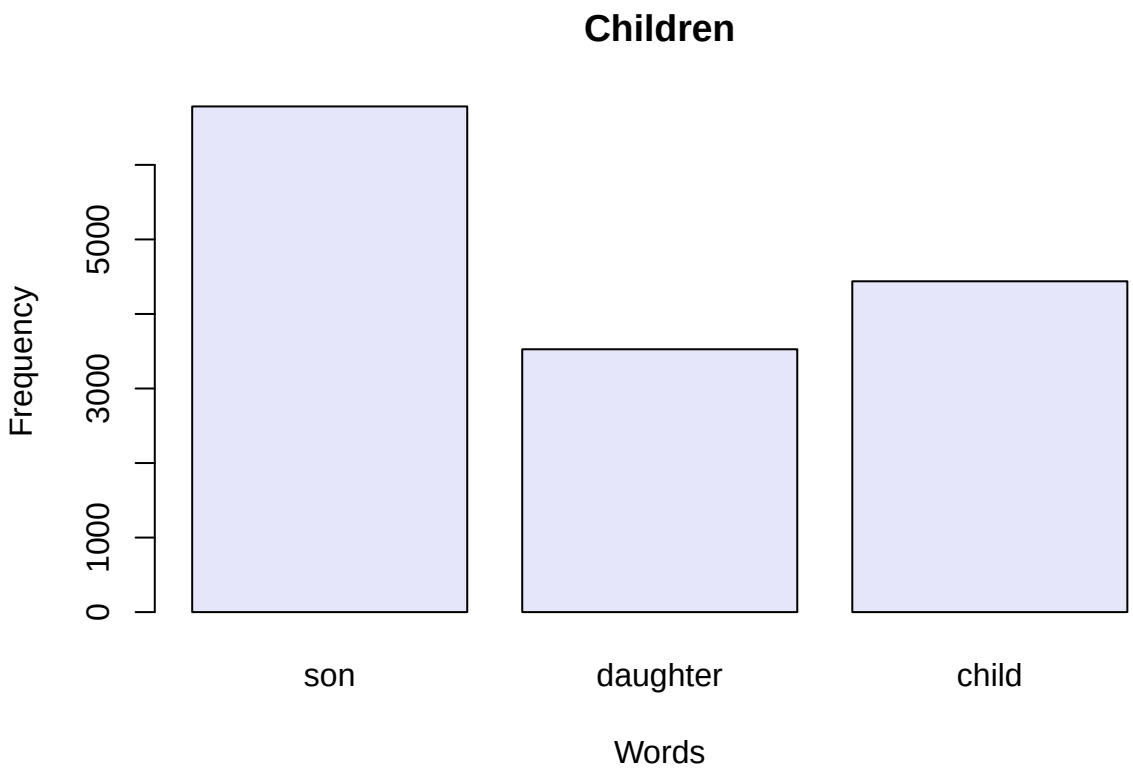

```r
siblings <- data.frame(
  words = c("brother", "sister", "sibling"),
  occurence = c(sum(brother), sum(sister), sum(sibling)))
str(siblings)
```

```
## 'data.frame':    3 obs. of  2 variables:
##  $ words    : chr  "brother" "sister" "sibling"
##  $ occurence: int  1494 1728 88
```

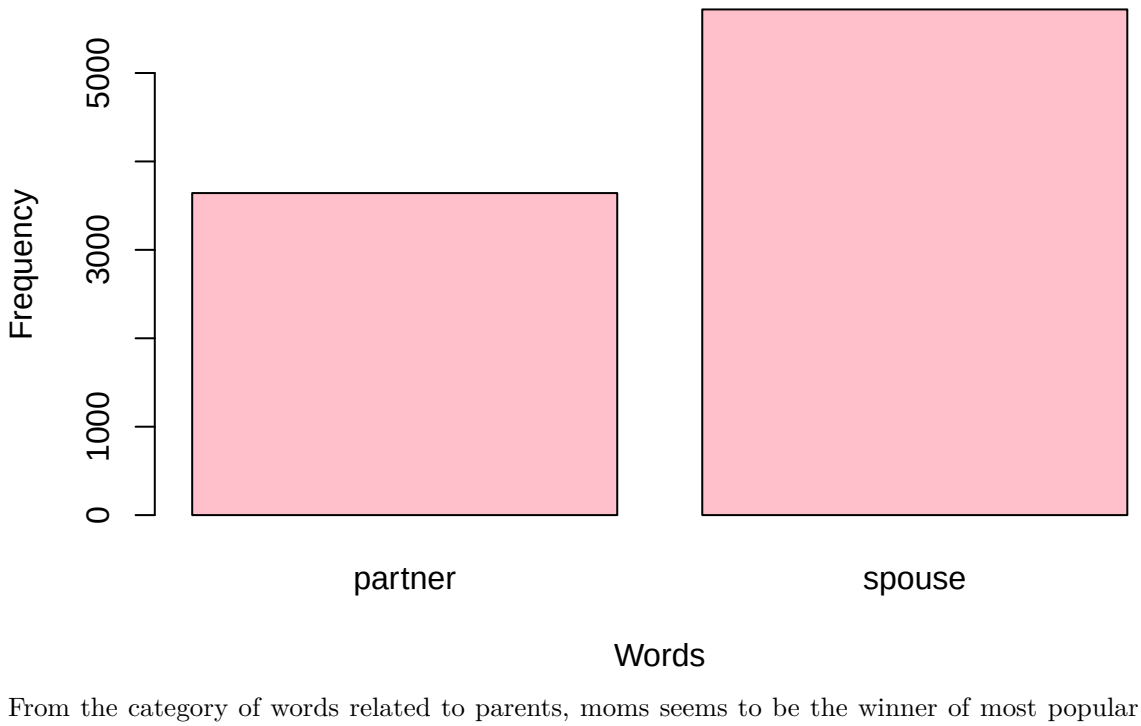
```r
barplot(height = siblings$occurence, name = siblings$words,
        xlab = "Words", ylab = "Frequency", main = "Siblings",
        col = "hotpink")
```




```
children <- data.frame(
  words = c("son", "daughter", "child"),
  occurence = c(sum(son), sum(daughter), sum(child)))
barplot(height = children$occurence, name = children$words,
        xlab = "Words", ylab = "Frequency", main = "Children",
        col = "lavender")
```




```
relationships <- data.frame(
  words = c("partner", "spouse"),
  occurence = c(sum(partner), sum(spouse)))
barplot(height = relationships$occurence, name = relationships$words,
        xlab = "Words", ylab = "Frequency", main = "Relationships",
        col = "pink")
```



From the category of words related to parents, moms seems to be the winner of most popular source of happiness. In the sibling category, sisters are the most popular mention. Out of all of those in the children category, the sons seem to be mentioned the most. Though there may be some overlap in terminology, those who are married seem to mention their spouse more than those who are not married, or refer to their partner in non-traditional married terms.

```
# pets
cat <- str_detect(words, "cat|Cat|kitten|Kitten|kitty|Kitty")
sum(cat)
```

```
## [1] 4199
```

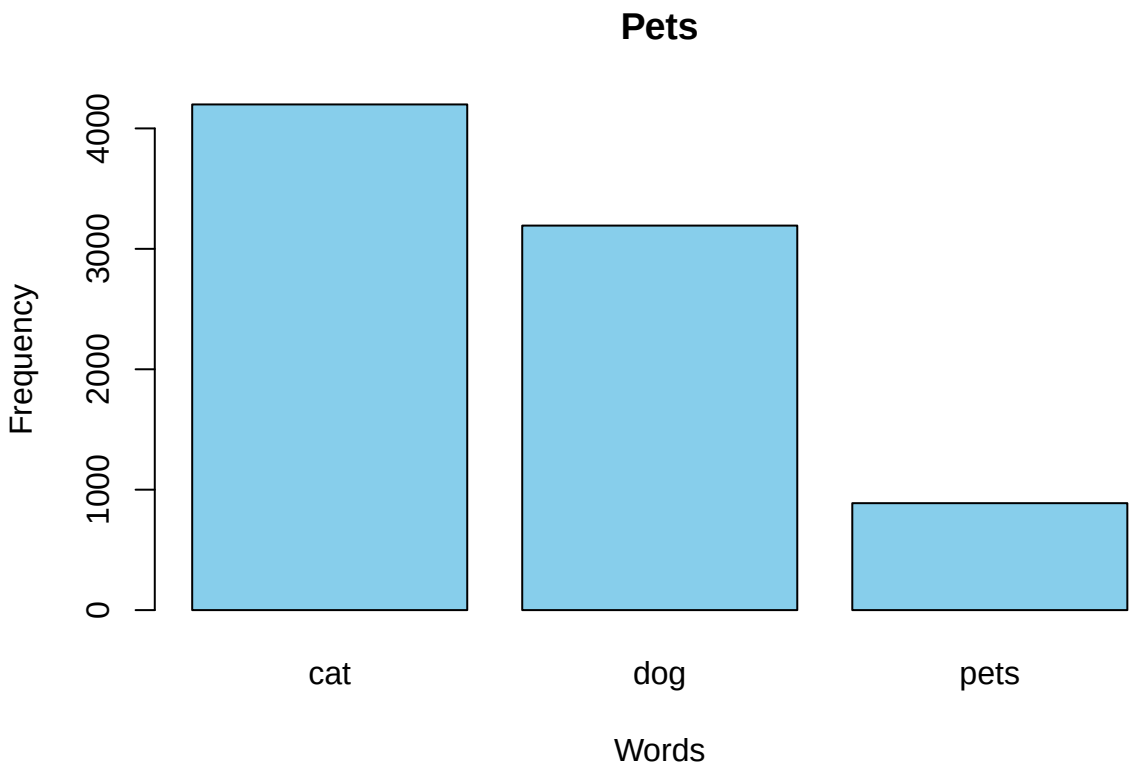
```
dog <- str_detect(words, "dog|Dog|puppy|Puppy|pup|Pup")
sum(dog)
```

```
## [1] 3193
```

```
pets <- str_detect(words, "pet+|Pet+")
sum(pets)
```

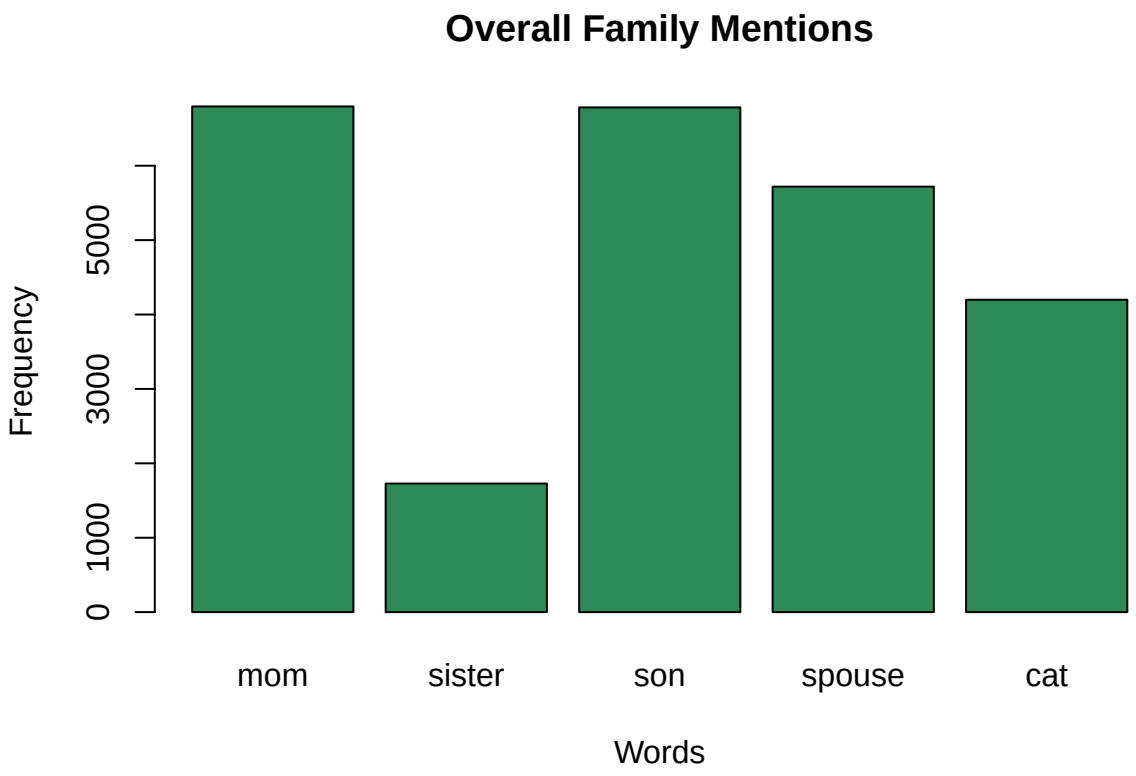
```
## [1] 888
```

```
pet <- data.frame(
  words = c("cat", "dog", "pets"),
  occurence = c(sum(cat), sum(dog), sum(pets)))
barplot(height = pet$occurence, name = pet$words,
        xlab = "Words", ylab = "Frequency", main = "Pets",
        col = "skyblue")
```



Finally, out of the pet category, cats are mentioned the most!

```
overall_family <- data.frame(
  words = c("mom", "sister", "son", "spouse", "cat" ),
  occurence = c(sum(mom), sum(sister), sum(son), sum(spouse), sum(cat)))
barplot(height = overall_family$occurence, name = overall_family$words,
        xlab = "Words", ylab = "Frequency", main = "Overall Family Mentions",
        col = "seagreen")
```



Though the competition between mentions of moms and sons is quite close, moms slightly edge out the sons with 6797 mentions, while sons have 6784 mentions.

A subexploration on food

This is not a listed category, but when I think about one of my greatest sources of happiness, I cannot help but think of food! Here are the mentions of words that would indicate the essence of a food related response:

```
food <- str_detect(words, "food|Food")
sum(food)
```

```
## [1] 1655
```

```
meal <- str_detect(words, "snack|Snack|breakfast|Breakfast|Lunch|lunch|dinner|Dinner")
sum(meal)
```

```
## [1] 6060
```

```
eat <- str_detect(words, "eat|Eat|ate|Ate|eating|Eating|eaten|Eaten")
sum(eat)
```

```
## [1] 18461
```

```
delicious <- str_detect(words, "delicious|Delicious")
sum(delicious)
```

```
## [1] 1114
```

```
yummy <- str_detect(words, "yummy|Yummy|yum|Yum")
sum(yummy)
```

```
## [1] 90
```

```
scrumptious <- str_detect(words, "scrumptious|Scrumptious")
sum(scrumptious)
```

```
## [1] 4
```

```
wonderful <- str_detect(words, "wonderful|Wonderful")
sum(wonderful)
```

```
## [1] 558
```

```
favorite <- str_detect(words, "favorite|Favorite")
sum(favorite)
```

```
## [1] 2897
```

Of all these words, the word "eat" takes the prize of most populat mention!

Note: Limitations on food -

As aforementioned, food is a great contributor of joy to myself and many of the people in my life! However, this method of search with this broad of a dataset will not be able to tell the whole picture of if and how food is discussed in this dataset. One big reason is that it is quite likely that foods may be listed by name and it would be truly EXHAUSTING to use this method to search every possible food that could be named. Another reason why this topic poses an issue has to do with adjectives. There are many wonderful adjectives that could be used to describe how or why food brings someone joy. Although a few are mentioned above, there are a great number of adjectives that could be used to describe food that could also be used in general to describe anything else that brings happiness (wonderful, great, good, etc.)

A simple way to begin approaching this issue would be to create a subset of sentences that all contain the word food at least once. A similarly motivated subset could also be created out of all the sentences that use a form of the word eat.

```
food_sentences_true <- str_detect(sentences, "food+|Food+")
food_sentences <- str_subset(sentences,"food+|Food+")
sum(food_sentences_true)
```

```
## [1] 1474
```

There are 1,474 sentences that mention food in this corpus!

Now in this food subset, let's search for some of the words again to see if there is a difference in how these words appear!

```
delicious_f <- str_detect(food_sentences, "delicious+|Delicious+")
sum(delicious_f)
```

```
## [1] 138
```

```
yummy_f <- str_detect(food_sentences, "yummy|Yummy|yum|Yum")
sum(yummy_f)
```

```
## [1] 13
```

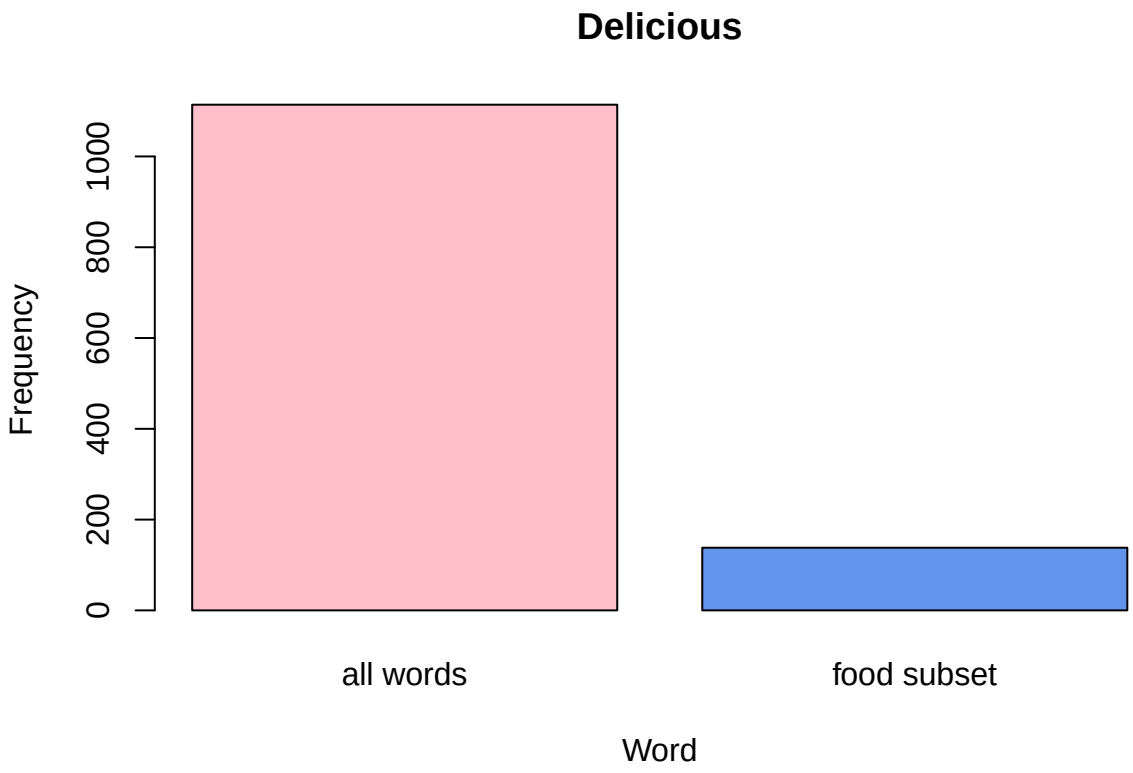
```
wonderful_f <- str_detect(food_sentences, "wonderful|Wonderful")
sum(wonderful)
```

```
## [1] 558
```

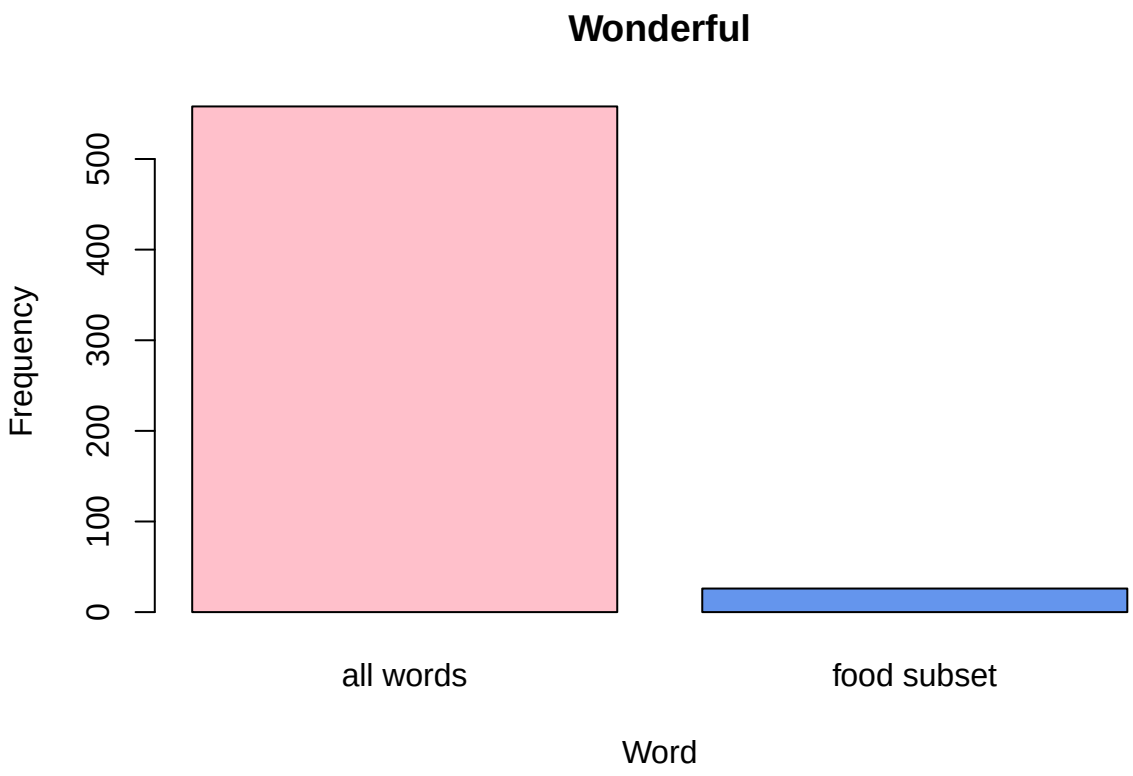
```
favorite_f <- str_detect(food_sentences, "favorite|Favorite")
sum(favorite_f)
```

```
## [1] 187
```

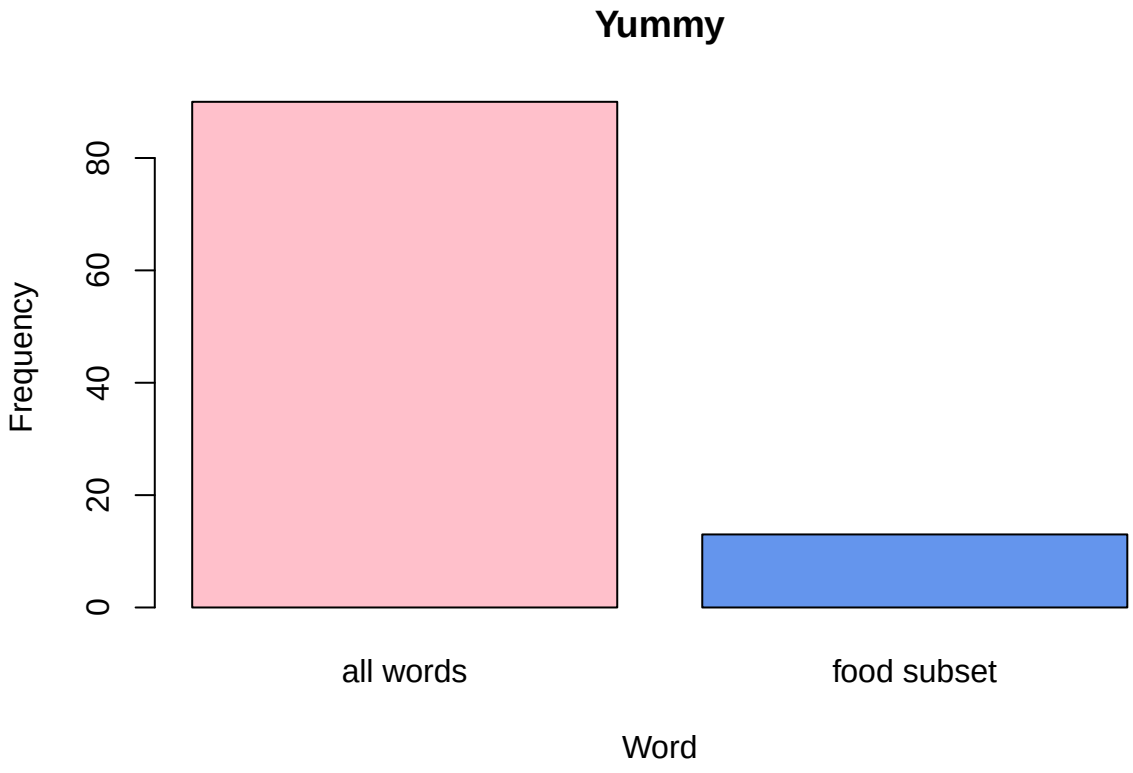
```r
# comparisons

delicious_total <- data.frame(words = c("all words", "food subset"),
                              occurence = c(sum(delicious), sum(delicious_f)))
barplot(height = delicious_total$occurence, name = delicious_total$words,
        xlab = "Word", ylab = "Frequency", main = "Delicious",
        col = c("pink", "cornflowerblue"))
```



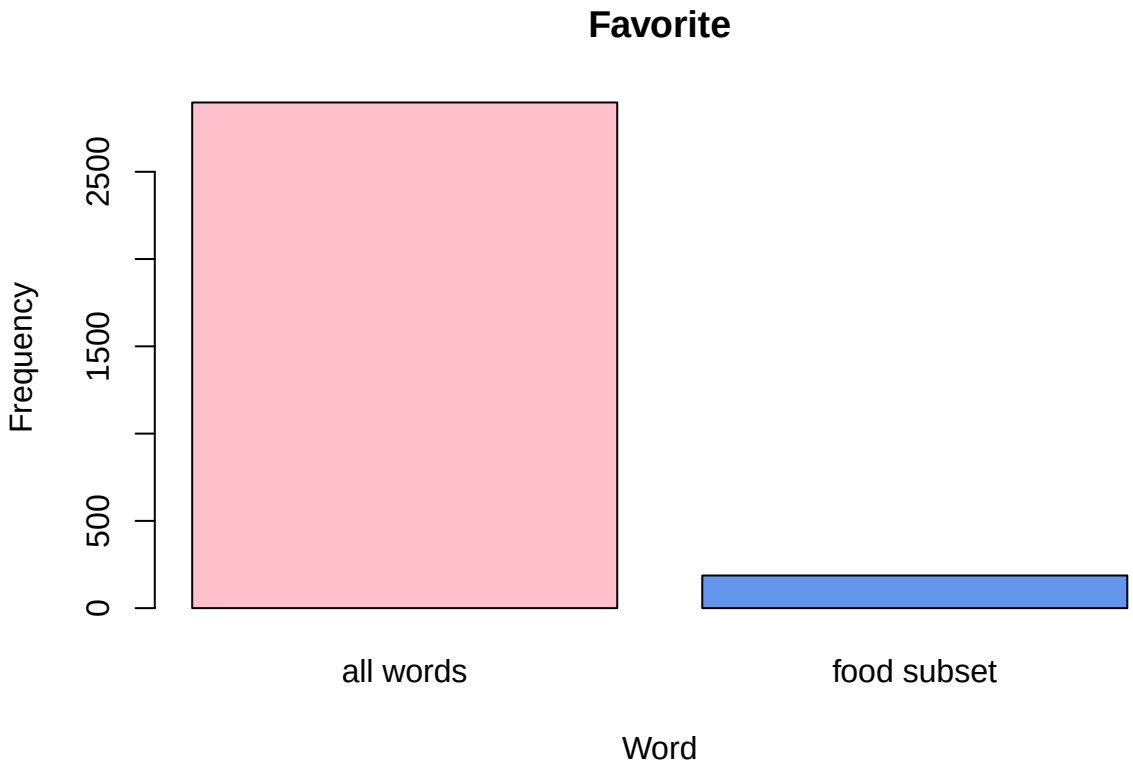

```r
wonderful_total <- data.frame(words = c("all words", "food subset"),
                              occurence = c(sum(wonderful), sum(wonderful_f)))
barplot(height = wonderful_total$occurence, name = wonderful_total$words,
        xlab = "Word", ylab = "Frequency", main = "Wonderful",
        col = c("pink", "cornflowerblue"))
```




```
yummy_total <- data.frame(words = c("all words", "food subset"),
                          occurence = c(sum(yummy), sum(yummy_f)))
barplot(height = yummy_total$occurence, name = yummy_total$words,
        xlab = "Word", ylab = "Frequency", main = "Yummy",
        col = c("pink", "cornflowerblue"))
```



```r
favorite_total <- data.frame(words = c("all words", "food subset"),
                             occurence = c(sum(favorite), sum(favorite_f)))
barplot(height = favorite_total$occurence, name = favorite_total$words,
        xlab = "Word", ylab = "Frequency", main = "Favorite",
        col = c("pink", "cornflowerblue"))
```



The great disparity present between the use of the same words in a subset of data specifically regarding the word food versus the use of the given adjective taken from the whole data set highlights and confirms the idea that greater specificity in the dataset will allow for clearer conclusions to be drawn (as there is a great potential for contextual confusion otherwise)! One other reason that is more likely a smaller contributor to the difference is that the subset considered the whole sentence rather than the words within the subset individually.