What are the Differences of Happiness Based on Gender?

Shiwei Hua - sh3804

HappyDB is a corpus of 100,000 crowd-sourced happy moments via Amazon's Mechanical Turk. You can read more about it on https://arxiv.org/abs/1801.07746 (https://arxiv.org/abs/1801.07746)

In this R notebook, we process the raw textual data for our data analysis.

I. Exploratory Data Analysis

1.0 - Load all the required libraries

From the packages' descriptions:

- `tm` is a framework for text mining applications within R;
- `tidyverse` is an opinionated collection of R packages designed for data science. All packages share an underlying design philosophy, grammar, and data structures;
- `tidytext` allows text mining using 'dplyr', 'ggplot2', and other tidy tools;
- `DT` provides an R interface to the JavaScript library DataTables.

1.1 - Load the data to be cleaned and processed

1.2 - Preliminary cleaning of text

We clean the text by converting all the letters to the lower case, and removing punctuation, numbers, empty words and extra white space.

1.3 - Stemming words and converting tm object to tidy object

Stemming reduces a word to its word *stem*. We stem the words here and then convert the "tm" object to a "tidy" object for much faster processing.

1.4 - Creating tidy format of the dictionary to be used for completing stems

We also need a dictionary to look up the words corresponding to the stems.

1.5 - Removing stopwords that don't hold any significant information for our data set

We remove stopwords provided by the "tidytext" package and also add custom stopwords in context of our data. In addition to the words provided in the start code, we add some other words such as "day", "time", "days", etc to the stopwords.

1.6 - Combining stems and dictionary into the same tibble

Here we combine the stems and the dictionary into the same "tidy" object.

1.7 - Stem completion

Lastly, we complete the stems by picking the corresponding word with the highest frequency.

1.8 - Pasting stem completed individual words into their respective happy moments

We want our processed words to resemble the structure of the original happy moments. So we paste the words together to form happy moments.

1.9 - Keeping a track of the happy moments with their own ID

Exporting the processed text data into a CSV file

Combine both the data sets and keep the required columns for analysis

We select a subset of the data that satisfies specific row conditions.

Here, after looking to the data again, we found that there are some observations does not make sense. For example, there are 9 people whoes age is 227. And also, for people younger than 3 years old, we don't think they can write these notes. Therefore, we clean the data again.

II. Start to our research question: What are the Differences of Happiness Based on Gender?

1. Overall Word Cloud

According to the overall Word-Cloud, we found that “frien”d is a main factor of happiness among all people, also, we can see that “family”, “played”, “home”, etc are also important factors for one to be happy. In this project, I will specificlly look at how happiness differ between males and females.

2. Word Cloud based on gender

2.1 Word Cloud for Males

According the word cloud above, we see that friend is the most important one as expected. However, other things such as “wife”, “family”, “game”, “played” also showed their weight on males’ happiness.

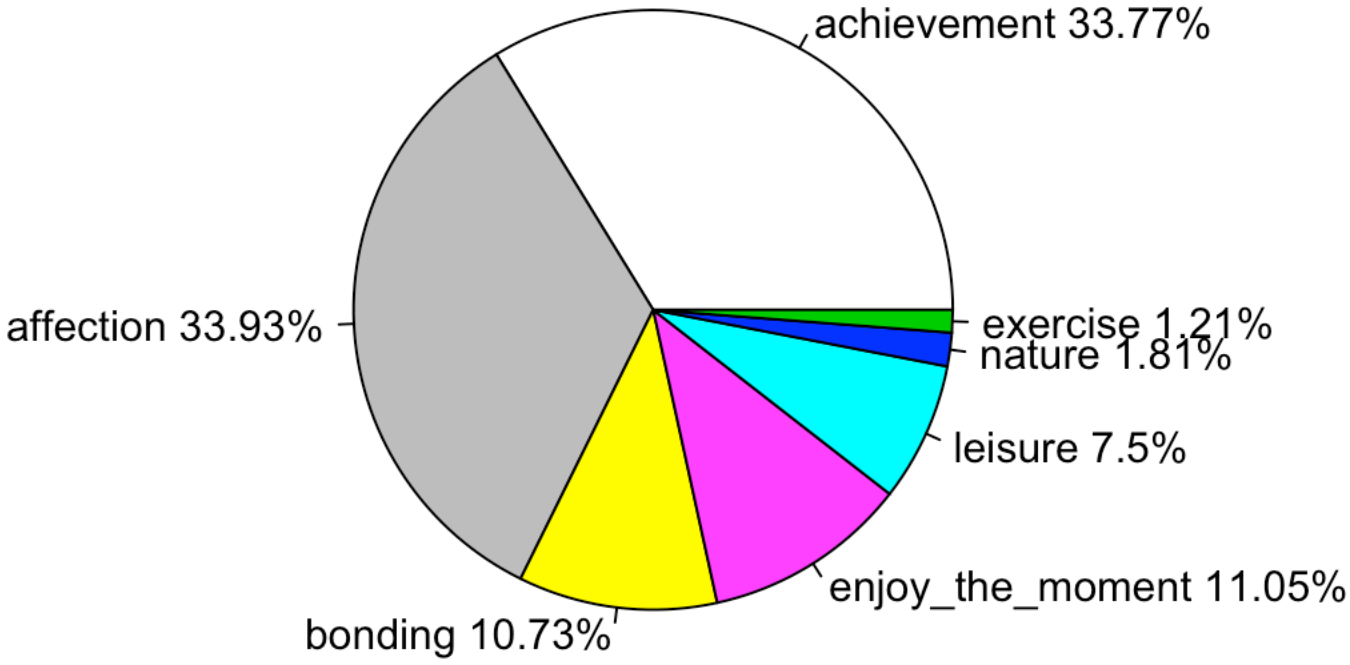
2.2 Word Cloud for Females

For females, besides “friends”, there are “husband”, “home”, “family”, “son”, “daughter” as main effects of females’ happiness. This may shows that most of females get their happiness from their family aspects.

3. Analysis of Happiness Catagories

3.1 Pie chart for the happiness catagory

After the glimpse on the word clouds, we have a general sense about what made people happy. They, we study their happiness in to more details; specifically, through seven categories. They are “achievement”, “affection”, “bonding”, “enjoy_the_moment”,“leisure”,“nature”,and “exercise”.

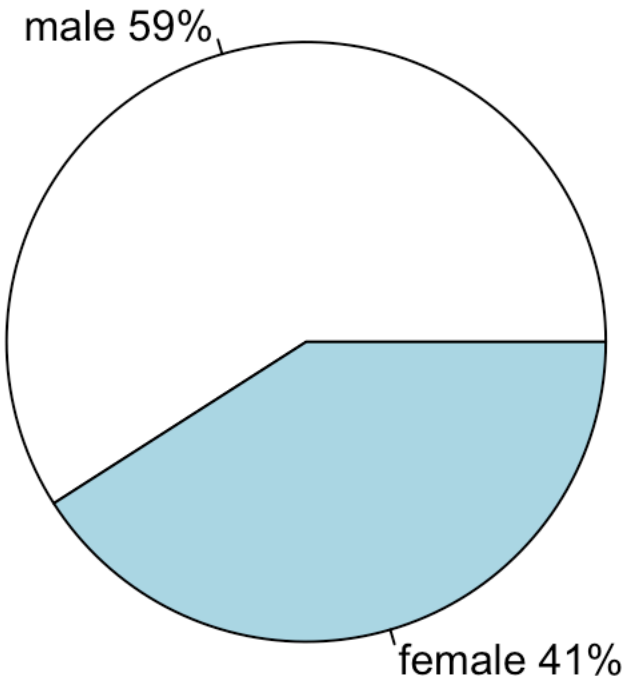


After we look at the pie chart above, we noticed that for all people, affections and achievement and even bonding are major contributors to one’s happiness. Again, we will look at the differences of happiness categories based on gender.

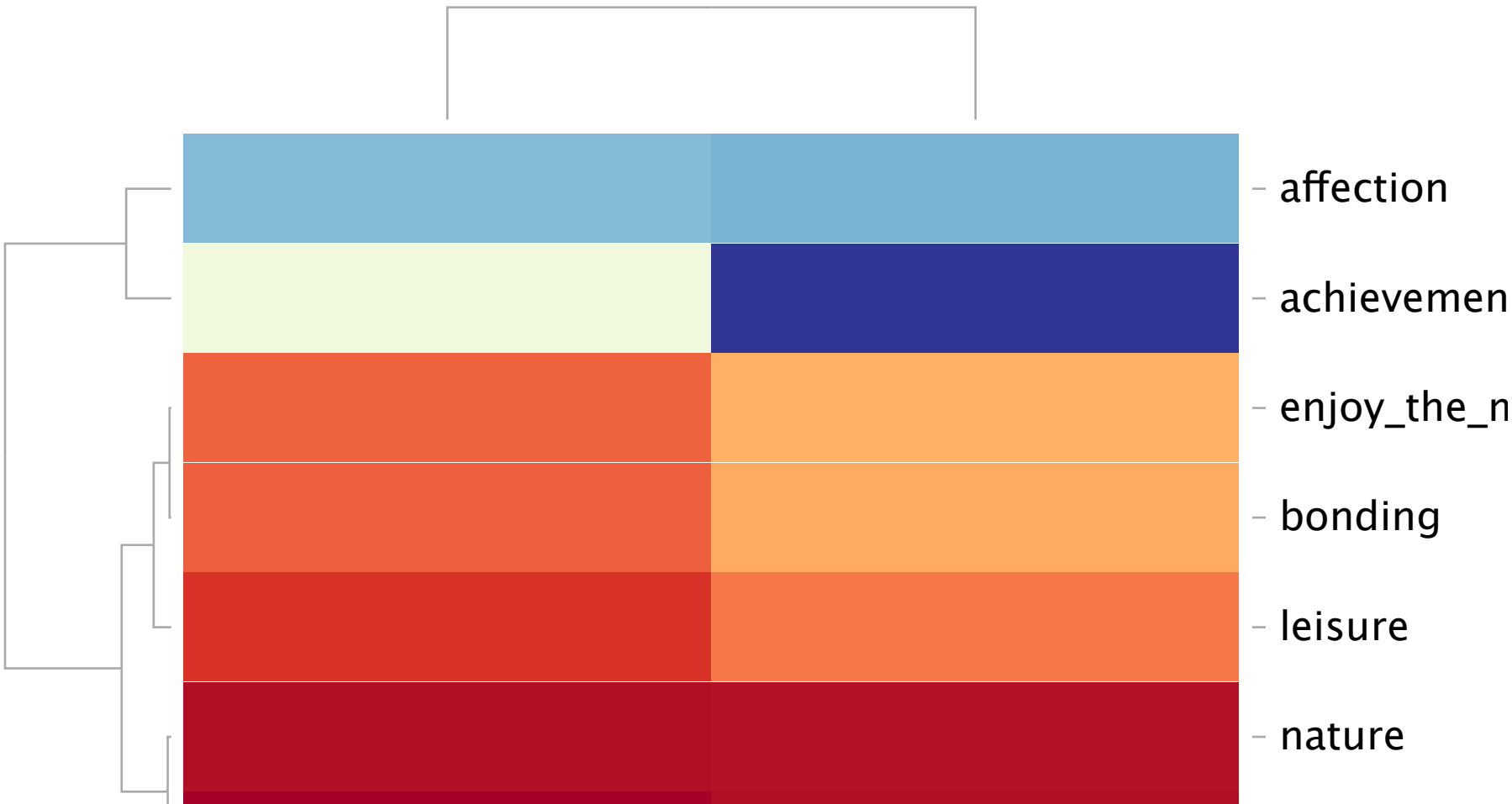
3.2 happiness catagory based on gender

Here, we will use Heatmaps to study these behaviors. Since Heatmaps are based on the word counts, we need to see whether the number of males equals to the number of females in our data. Thus, we create a pie chart to visualize this.

Where are the data from?



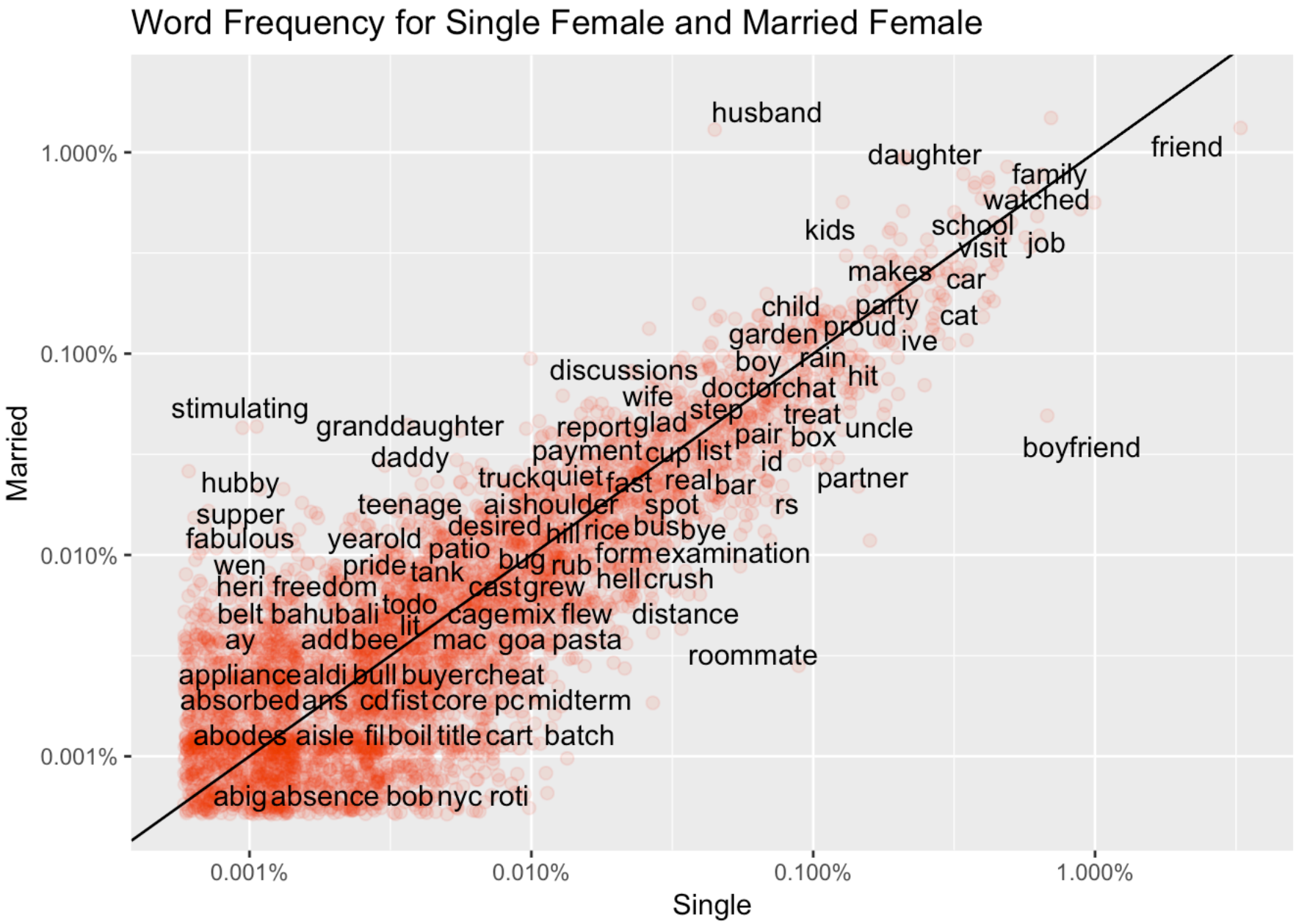
According to the pie chart above, we noticed that in our data, 59% are from males and 41% are from females, which means, this data for gender are not even. We needs to take special consideration to this point in the following studies. Then, we create a heatmap.



According to the heatmap above, we noticed that there is a huge difference in achievement and exercise between males and females. It seems that males consider more about their happiness through achievement and exercises. Then, we noticed that for affections, the number are almost the same bewteen males and females. However, as we mentioned above, in our data, the ratio of male to female is approximate 6:4. Thus, this means that more precentage of females feels happy through affections than that of males.

4. Word frequency plot to check whether **marital** is influencial in the happiness

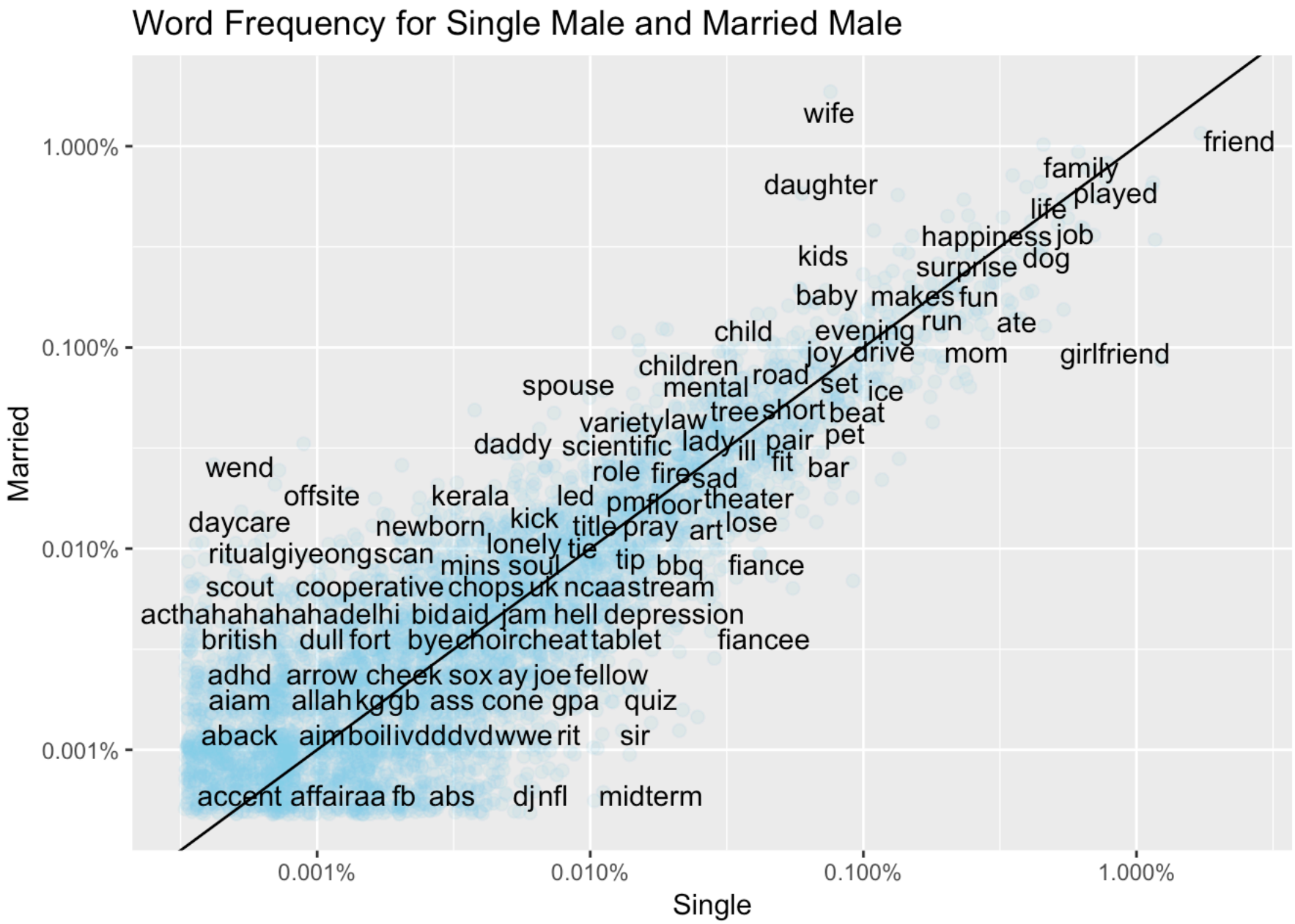
4.1 Word frequency plot to check whether **marital** is influencial to **Female**



According to the word frequncy plot, we finds out that Marriage is influencial in females' happiness. We observed that among married females, "husband", "daughter", "kids" has a relatively high frequnecy. While among single females, "boyfriend", "partner", "roommate" has a relatively high frequnecy. This make sense

because when people married, they care more about their family. When people is single, they care more about their relationships and surrounding friends.

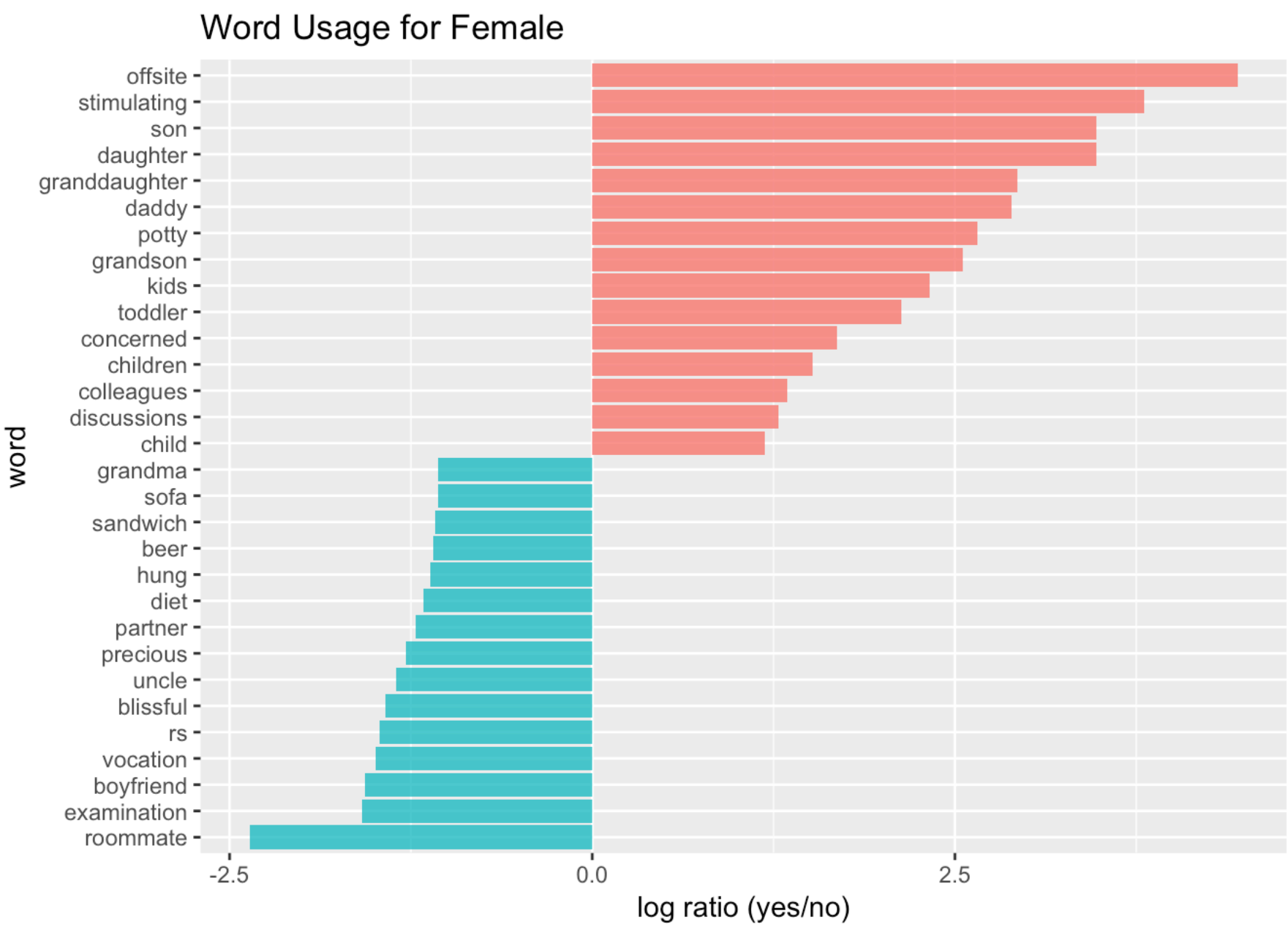
4.2 Word frequency plot to check whether **marital** is influencial to **Male**



We saw a similar pattern of Males. Married Males cares more about their family, that is why words such as “wife”, “daughter”, “kids”, etc has more weight. While for single Males, of course “girlfriend” has large weight. But beside that, usually they are relatively young, so they care more about themselves and their futures. That is why “internship”, “quiz”, “midterm” also have a relatively high frequncy.

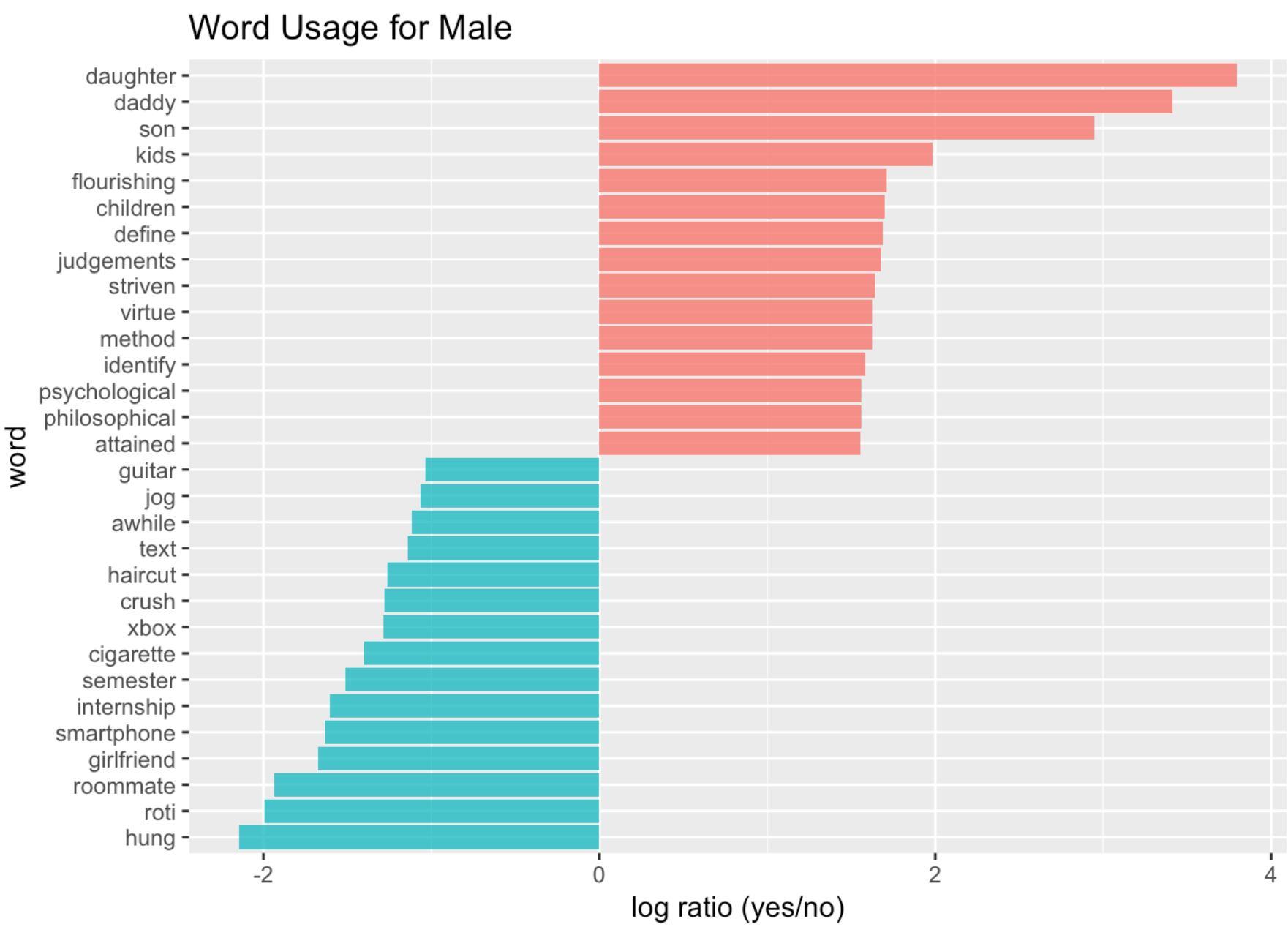
5. Whether **Parenthood** influence the happiness?

5.1 **Parenthood** of Female



After look at the bar plot above, we finds out that Parenthood is also influencial in Females' happiness. After becoming a parent, they care more about there children, which makes "son", "daughter", "granddaughter" etc appears more frequently compared with those who is not a parent. While for females not becoming a parent, most of them may be single; that is why "roommate", "boyfriend", "vocation" has more weights.

5.2 **Parenthood** of Male



Also, we find out that Parenthood is influencial in Males' happiness. After becoming a parent, similar to females, they care more about there children, which makes "son", "daughter", "kids" etc appears more frequently compared with those who is not a parent. While for males not becoming a parent, most of them may care less about their family; that is why "roommate", "girlfriend", "hung" has more weights.

III. Summary

1. Overall, **Friendship** is the most popular effect to make people happy. For females, family would be a main effect and for males, beside family, their own entertainments such as games are also important aspect to make them happy.
2. The sources of happiness for females comes more from affections and that of males comes more from achievements and exercise.
3. People's marital and parenthood also influences the reasons that makes them happy, and the influence of marital and parenthood have high correlations. After marriage, people cares more about their offsprings and family; While before marriage, females care more about their relationships and surrounding friends, while males care more about their future and career/study performance.

Reference

1. https://github.com/rit-public/HappyDB (https://github.com/rit-public/HappyDB)