Project1 What made you happy today?

Nannan Wang (nw2387)

1.Data Preparation

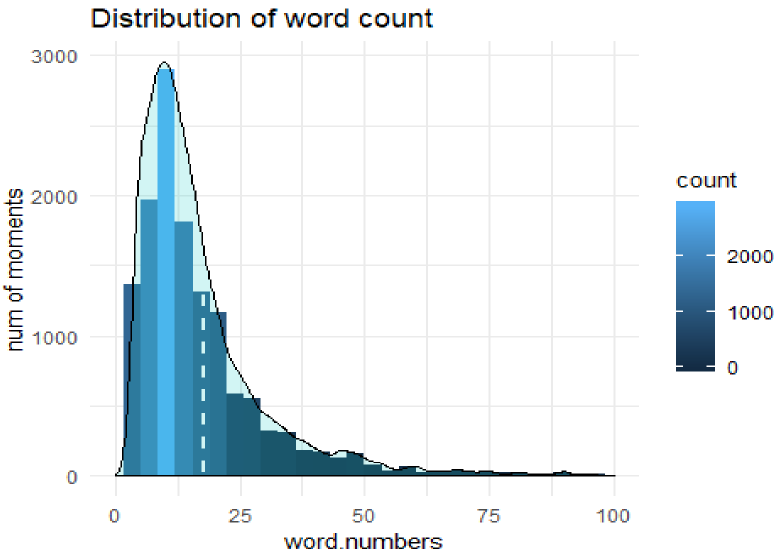
In this part, we prepare the Happy Moment datasets given on the website.
We need to combine the 'demographic.cvs' with 'cleaned_hm.cvs' and then omit the NA data.

2.Data Presentation

In this part, we use different forms to display the happy moments' text and explore some interesting details.

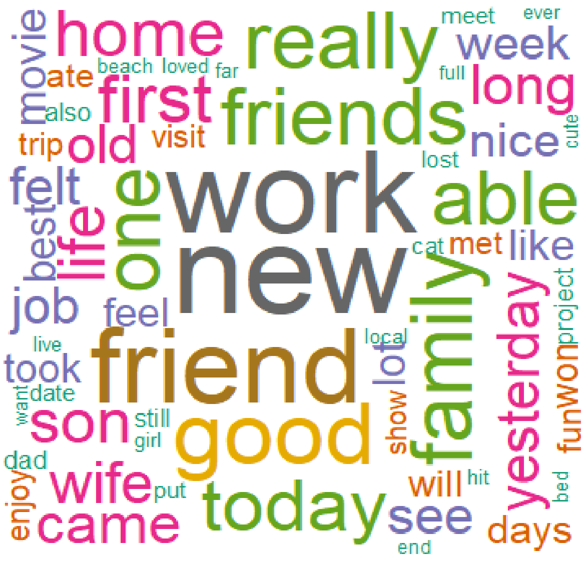
2.1 Word Count

In this Part, I count the word number for each happy moment discribtion, and most people can express their happiness with less than 15 words. It perhaps shows that happiness do not need too much words to speak out.



2.2 Word Frequency

In this part, we find that some words appear most frenquently in people's happy moments, such like: "work", "friend", "new", "family", "son","game", "birthday" etc.



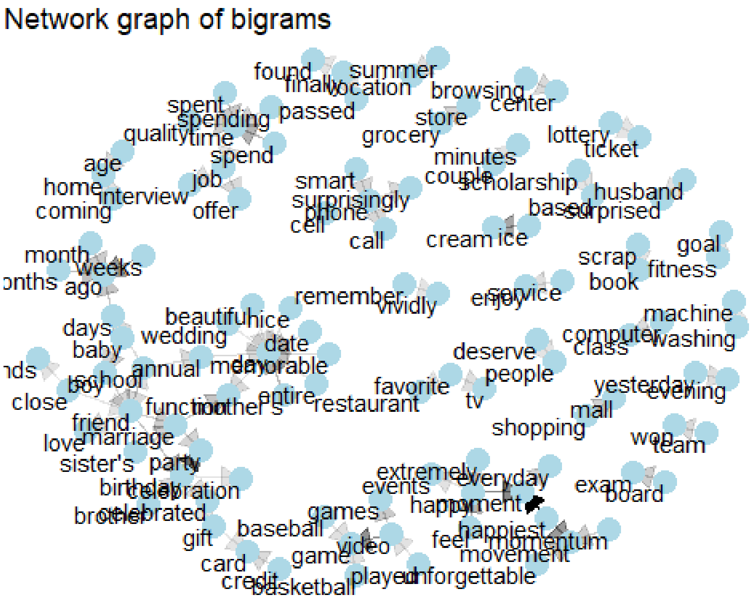
2.3 Bigrams

```
## # A tibble: 10 x 3
##    word1     word2        n
##    <chr>     <chr>    <int>
##  1 happiest moment     191
##  2 birthday party      105
##  3 happy    moment      86
##  4 video    game        79
##  5 happiest movement    73
##  6 weeks    ago         73
##  7 3        months      66
##  8 ice      cream       63
##  9 24       hours       55
## 10 feel     happy       53
```

In this part, we focus on the bigrams which are phrases we used in the daily life. In terms of top 10 bigrams, we find top three meaningful phrases which play very important roles in people's happy moments:1)birthday party 2)video game 3)ice cream, that is amazing!

2.4 Bigrams Visualization

In this part, based on the bigrams we focused, we visualize the bigrams using the network graph, and we can find the most popular happy moments and their connections! For instance, the wedding/date/marriage can be connected together.

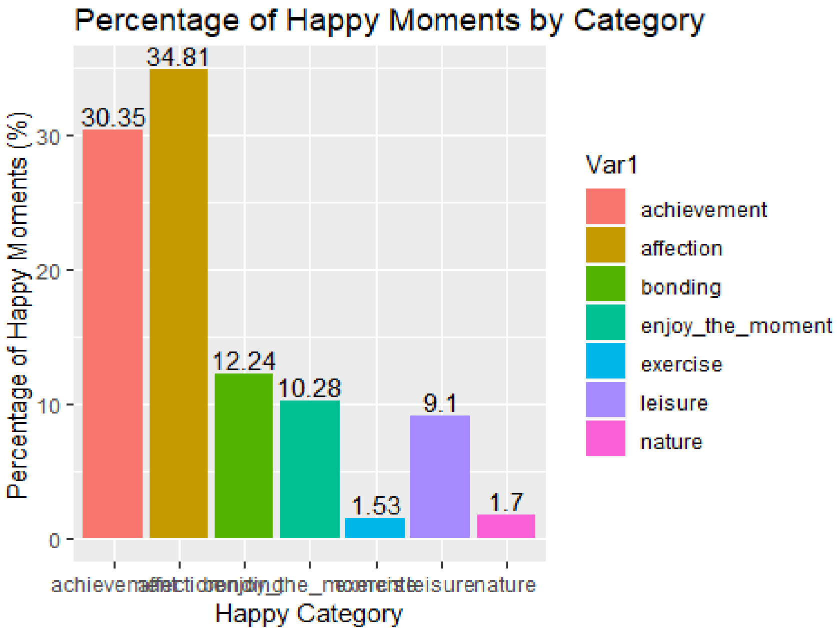


3.Exploration of Happy Moments

In this part, we wan to explore different Categories of Happy Moments, and we also explore the moments by different age groups/ gender groups/ marital groups.

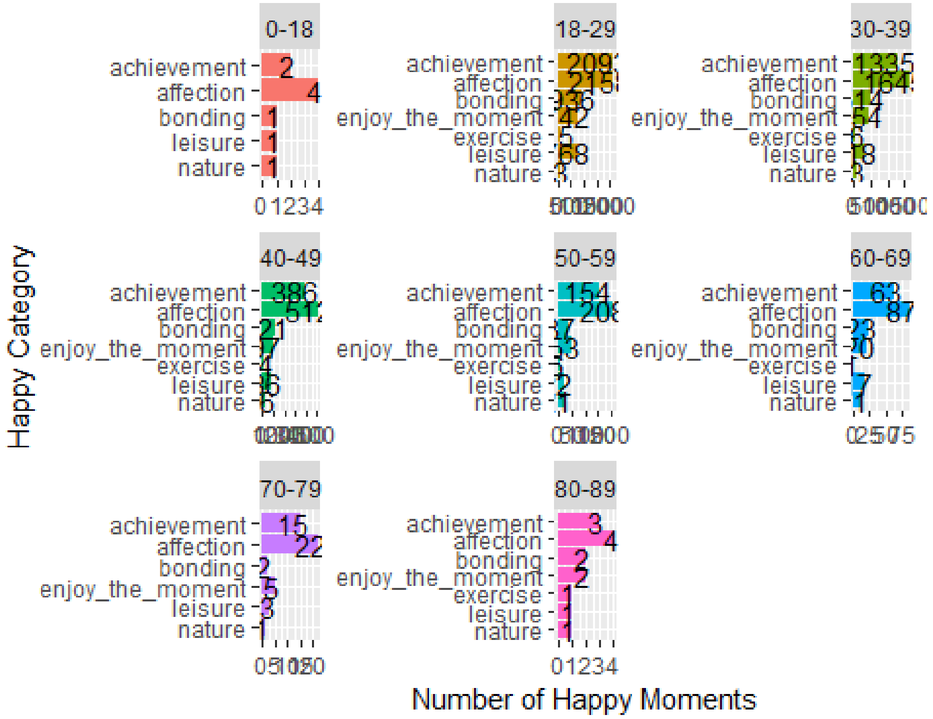
3.1 Percentage of Happy Moments

In this part, we find that based on all the happy moments, people get their happiness from affenction most, and then the achivement, which is related to the wordcloud we generated above.



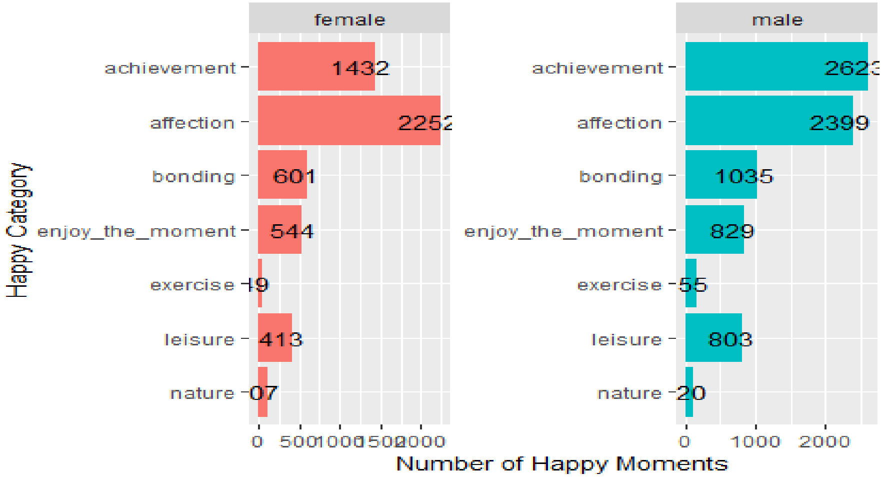
3.2 Happy Category by Age

In this part, based on different age group, we can not explore significant difference bewteen different groups, they all get happiness from affection most, and second is the achievement.



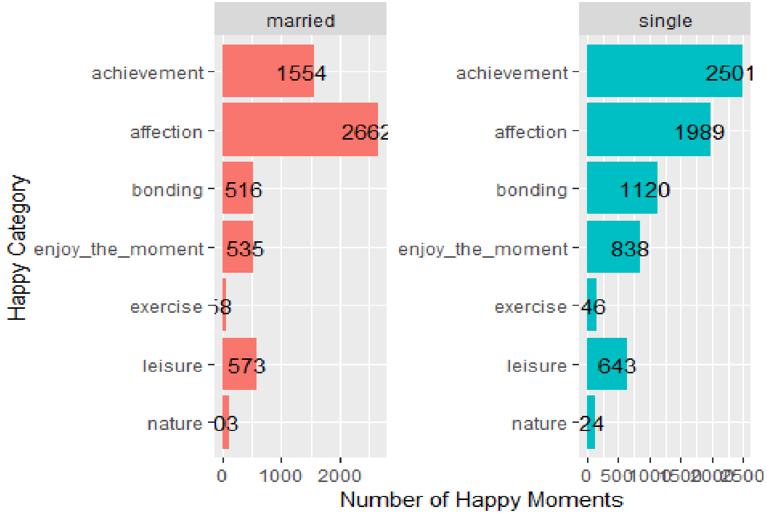
3.3 Happy Category by Gender

In this part, we find that male get happiness from achievement most which is totally different from the results above, and also exercise accounts for a lot by male than female. That makes sense!



3.4 Happy Category by Marital Status

In this part, we find that single people get happiness from achievement most and the affection is the second, which perhaps means that single people have less happiness from affection whithout their own kids and husband(wife).

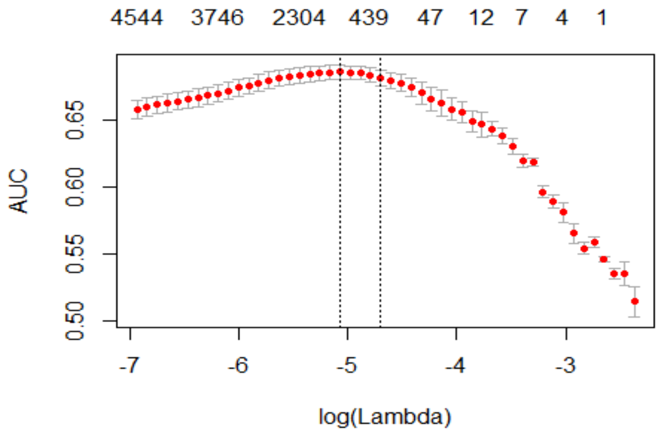


4. Logistic Regrssion

In this part, we want to explore deeper in people's happy moments, we can apply logistic regression to build some classifiers to recognize people's gender/marital status/parenthood status according to their happy moment descriptions.

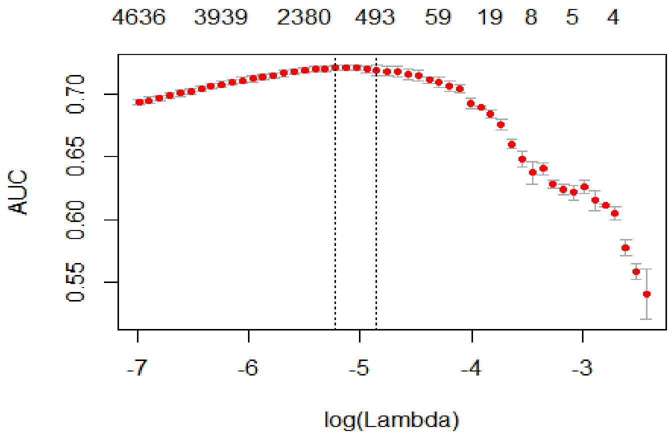
4.1 Classifier for Gender

In this part, we make a classifier to classify people's gender. First, split 70% data randomly as train data and the rest 30% are test data. And then apply the logistic regression for train data. At last test the calssifier on the test data. In this case, we find the the classifier accurate is about 0.7297626 tested on the test data, the classifier works pretty well.



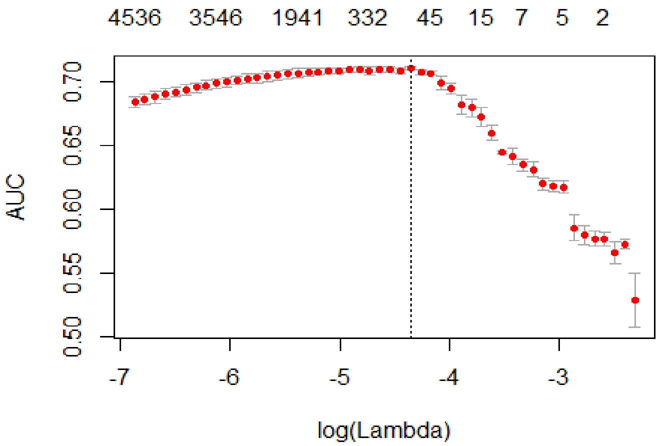
4.2 Classifier for Marital Status

The same method as 4.1, and the accurate of marital classifier is 0.7783142, it also works well. The graph shows the max AUC is 0.7209.



4.3 Classifier for Parenthood

The same method as 4.1, and the accurate of parent classifier is 0.7241155, it also works well.


AUC
log(Lambda)