Lab report

Load data

```
wordbank <- read_csv("https://dyurovsky.github.io/85309/data/lab3/wordbank.csv")
```

```
## Rows: 1020000 Columns: 6
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (3): gender, category, word
## dbl (2): id, age
## lgl (1): knows
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Exercise 1:

there are 677 words and 22 categories in the form. the category with the most words is action_words with 103 words in it.

```
# enter your code for Exercise 1 here
words_count <- wordbank %>%
  distinct(word)%>%
  summarise(n=n())

categories_count <- wordbank %>%
  distinct(category)%>%
  summarise(n=n())


categories_most_words <- wordbank %>%
  group_by(category) %>%
  distinct(word) %>%
  summarise(count = n()) %>%
  arrange(desc(count))

words_count
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1   677
```

```
categories_count
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1    22
```

```
categories_most_words
```

```
## # A tibble: 22 x 2
##    category          count
##    <chr>             <int>
##  1 action_words        103
##  2 food_drink           67
##  3 descriptive_words    63
##  4 household            50
##  5 animals              43
##  6 furniture_rooms      33
##  7 outside              31
##  8 people               29
##  9 clothing             28
## 10 body_parts           27
## # ... with 12 more rows
```

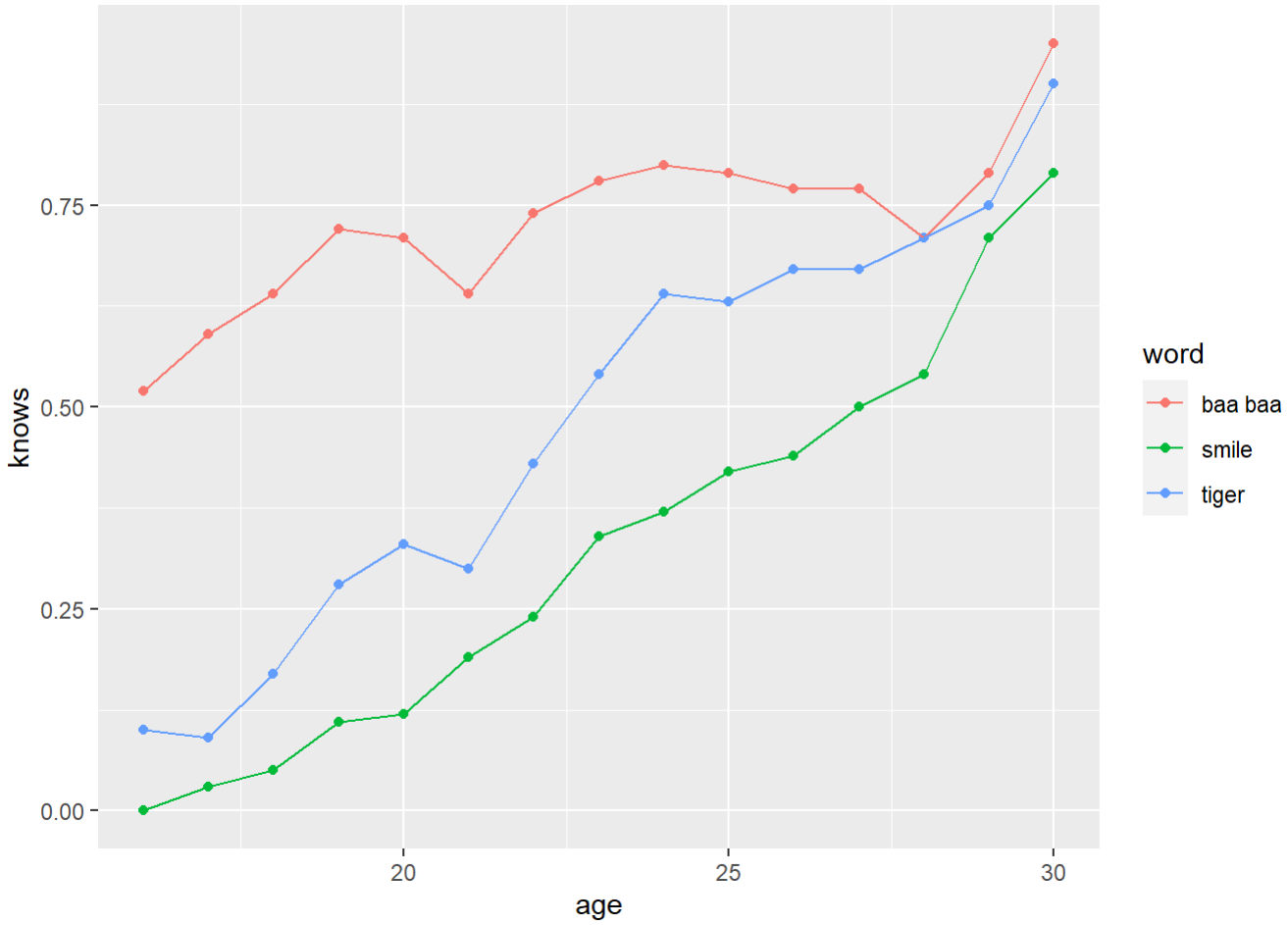
Exercise 2:

A word easier than tiger to learn is “baa baa”, since more people knows it in each age (specially in earlier ages) compared to “tiger” and “Smile”. And “Smile” is the hardest to learn compared to “tiger” since less kids know it in each age.

```
# enter your code for Exercise 2 here
words_selected <- wordbank %>%
  filter(word %in% c("tiger","baa baa","smile")) %>%
  group_by(age, word) %>%
  summarise(knows = mean(knows))
```

```
## `summarise()` has grouped output by 'age'. You can override using the `.groups` argument.
```

```
ggplot(words_selected, aes(x = age, y = knows, color = word)) +
  geom_point() +
  geom_line()
```



Exercise 3:

The hardest word is “country” because it has the lowest prob (or mean of kids who knows it at different ages), and the easiest words is “mommy” because it has the highest prob (or mean of kids who knows it at different ages). “mommy’ makes sense because moms are the closet ones to children, but I’d expect to see another word as the hardest than “country”. The reason might be that children at those ages are not familiar with the context of countries, etc.

```
# enter your code for Exercise 3 here
word_difficulty <- wordbank %>%
  group_by(word, age)%>%
  summarise(S = mean(knows)) %>%
  summarise(prop = mean(S)) %>%
  arrange(desc(prop))
```

```
## `summarise()` has grouped output by 'word'. You can override using the `.groups` argument.
```

```
word_difficulty
```

```
## # A tibble: 677 x 2
##    word   prop
##    <chr> <dbl>
##  1 mommy 0.972
##  2 daddy 0.968
##  3 ball  0.939
##  4 hi    0.911
##  5 bye   0.902
##  6 uh oh 0.899
##  7 no    0.888
##  8 dog   0.885
##  9 shoe  0.877
## 10 baby  0.867
## # ... with 667 more rows
```

Exercise 4:

the 19 months old child's hardest word is "sister" and the 30 months old child's hardest word is "snowsuit", which generally makes sense, because "snowsuit" is a harder word than "sister", and it is understandable that the 19 months old child des not now "snowsuit" while the 30 months old child does.

```
# try printing this out to see what left_join did
wordbank_difficulty <- left_join(wordbank, word_difficulty, by = "word")

hardest_word <- function(child) {
  word_tibble <- wordbank_difficulty %>%
    filter(id == child, knows == TRUE) %>%
    arrange(prop)
  # enter your code for Exercise 4 here and uncomment the pipe above

  word_tibble %>%
    pull(word) %>%
    first() # get the first word if there are multiple
}

hardest_word("129277")
```

```
## [1] "sister"
```

```
hardest_word("129579")
```

```
## [1] "snowsuit"
```

More practice:

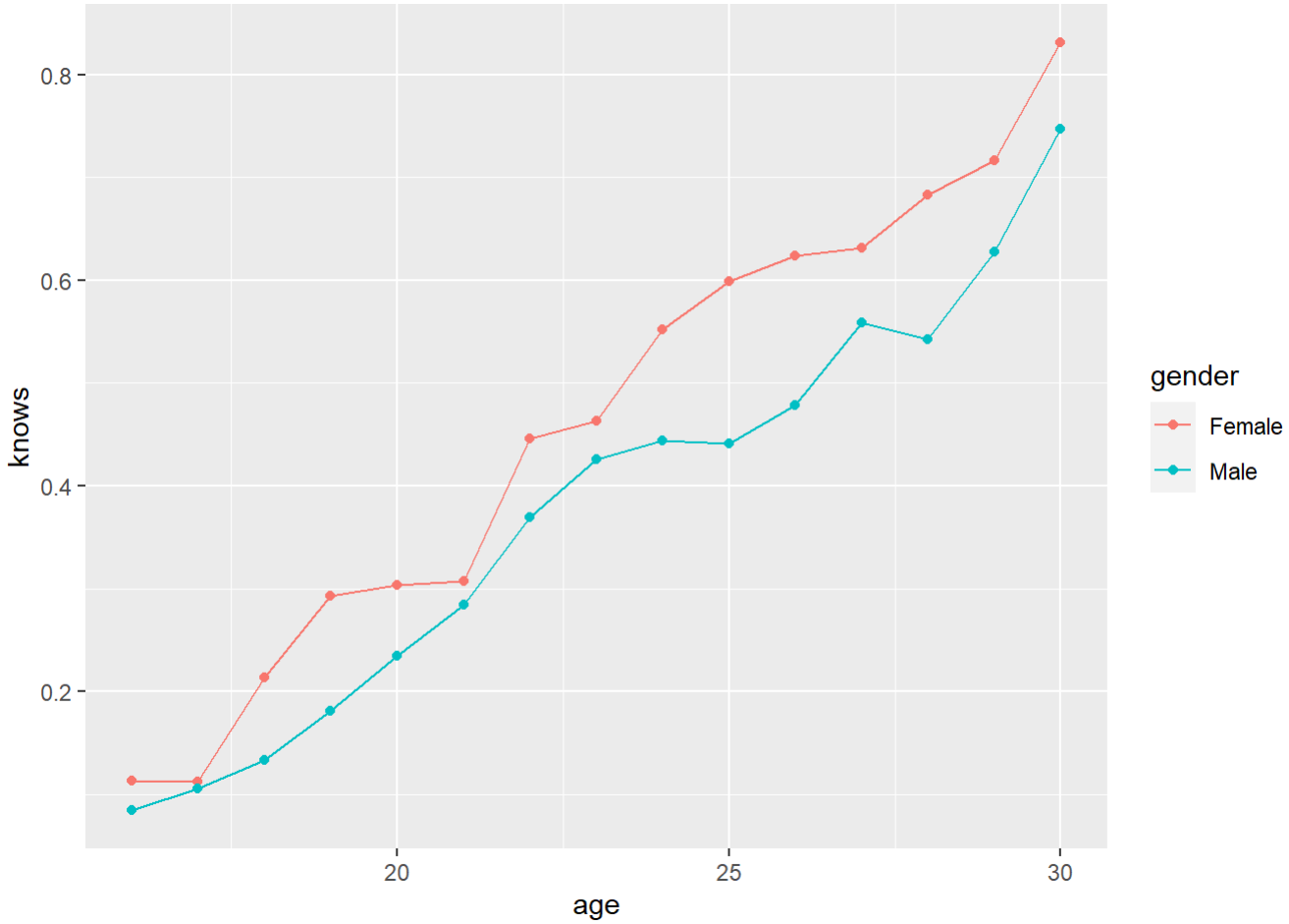
Exercise 5:

Yes, that is true, because at each age, more girls know more words than boys considering the higher rate of knows for girls than boys.

```
# enter your code for Exercise 5 here
words_selected <- wordbank %>%
  group_by(age, gender) %>%
  summarise(knows = mean(knows))
```

```
## `summarise()` has grouped output by 'age'. You can override using the `.groups` argument.
```

```
ggplot(words_selected, aes(x = age, y = knows, color = gender)) +
  geom_point() +
  geom_line()
```



Exercise 6:

Yes, there are children who knows all the words, and the youngest is 131265, who is a 25 months old kid. In this result, it is shown that there are children who know 680 words however based on the exercise 1, we have less than 680 words. The reason is that some words might be used in multiple categories.

```
# enter your code for Exercise 6 here
kids_count <- wordbank %>%
  group_by(id, age) %>%
  summarise(knowa = sum(knows))%>%
  ungroup()%>%
  filter(knowa == max(knowa)) %>%
  arrange(age)
```

```
## `summarise()` has grouped output by 'id'. You can override using the `.groups` argument.
```

```
kids_count
```

```
## # A tibble: 4 x 3
##       id   age knowa
##    <dbl> <dbl> <int>
## 1 131265    25   680
## 2 131161    26   680
## 3 129671    29   680
## 4 132054    30   680
```

Exercise 7:

The youngest child who knows “wish” is a 19 months old kid.

```
# enter your code for Exercise 7 here
youngest_age <- function(child_word) {
  word_tibble2 <- wordbank %>%
    filter(word == child_word, knows == TRUE) %>%
    arrange(age)
  # enter your code for Exercise 4 here and uncomment the pipe above

  word_tibble2 %>%
    pull(age) %>%
    first() # get the first word if there are multiple
}
youngest_age("wish")
```

```
## [1] 19
```