# OkCupid Lesson

*Kevin Winfree*

*3/3/2016*

## Text Processing, Mining, and Analysis of OkCupid User Profiles

## The Data

The we will be working with today consists of the public profiles of 59,946 OkCupid users who were living within 25 miles of San Francisco, had active profiles on June 26, 2012, were online in the previous year, and had at least one picture in their profile.

Steps to read in the data:
1. Download the folder to your computer.
2. Unzip the folder by double clicking the folder.
3. Re-upload the `.csv` file into your `Winfree_JSE_Project` folder.

### Lab: Are there differences between the genders in what words are used in the responses to the 10 essay questions?

```
profiles <- read.csv(file="profiles.csv", header=TRUE, stringsAsFactors=FALSE)
summary(profiles)
```

```
##       age          body_type             diet              drinks
##  Min.   : 18.00   Length:59946       Length:59946       Length:59946
##  1st Qu.: 26.00   Class :character   Class :character   Class :character
##  Median : 30.00   Mode  :character   Mode  :character   Mode  :character
##  Mean   : 32.34
##  3rd Qu.: 37.00
##  Max.   :110.00
##
##     drugs            education            essay0
##  Length:59946       Length:59946       Length:59946
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##     essay1             essay2             essay3
##  Length:59946       Length:59946       Length:59946
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
```

```
##     essay4             essay5             essay6
##  Length:59946       Length:59946       Length:59946
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##     essay7             essay8             essay9
##  Length:59946       Length:59946       Length:59946
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   ethnicity            height         income             job
##  Length:59946       Min.   : 1.0   Min.   :     -1   Length:59946
##  Class :character   1st Qu.:66.0   1st Qu.:     -1   Class :character
##  Mode  :character   Median :68.0   Median :     -1   Mode  :character
##                     Mean   :68.3   Mean   :  20033
##                     3rd Qu.:71.0   3rd Qu.:     -1
##                     Max.   :95.0   Max.   :1000000
##                     NA's   :3
##  last_online          location          offspring
##  Length:59946       Length:59946       Length:59946
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  orientation            pets             religion
##  Length:59946       Length:59946       Length:59946
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##      sex                sign              smokes
##  Length:59946       Length:59946       Length:59946
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##     speaks             status
##  Length:59946       Length:59946
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
```

```
##
##
```

```
#Packages and commands within the packages that we will be using in this Lab.

library(dplyr)      #commands associated with this package: select()
library(stringr)    #commands associated with this package: str_replace_all(), str_detect(), str_split(
library(mosaic)     #commands associated with this package: tally(), mosaicplot(), prop.test()
```

This chunk of code selects all of the essay columns, strings them together into a single vector and then each users vector of essay responses are separated by a space.

```
essays <- select(profiles, starts_with("essay"))            #select the 10 essay columns as identified
essays <- apply(essays, MARGIN=1, FUN=paste, collapse=" ")  #x is a user’s set of 10 essay responses an
essays <- str_replace_all(essays, "\n", " ")                #replace all line breaks with a space
essays <- str_replace_all(essays, "<br />", " ")            #replace all paragraph breaks with a space
```

**Determine in what percentage of male and female profiles a certain word appears.**

This chunk uses `str_detect()` from the `stringr` package to find how many times the word *book* appears in male and female essays and displays the proportions using `tally()`.

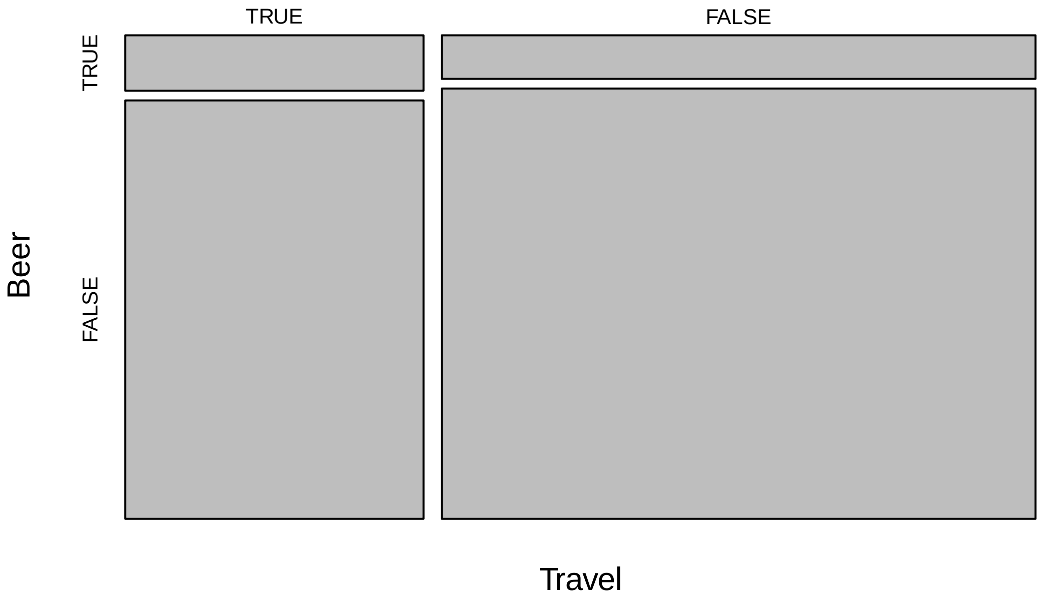
```
profiles$Has.Book <- str_detect(essays, "book")
tally(Has.Book ~ sex, profiles, format='proportion')
```

```
##          sex
## Has.Book          f         m
##     TRUE  0.6160385 0.5549694
##     FALSE 0.3839615 0.4450306
```

**Your Turn:** This time choose a word that is of interest to you and see how many time the word appears in the male and female essays. Display your results in proportions.

###Compare two words and give a visual representation of how often they occur in user profiles. Two interests of mine are “beer” and “travel”. Using a `mosaicplot` I would like to see a visual comparison of how often “beer” and “travel” appear in profiles together and separately.

```
profiles$Has.Beer <- str_detect(essays, "beer")
profiles$Has.Travel <- str_detect(essays, "travel")
Travel.vs.Beer <- tally(~Has.Travel + Has.Beer, data=profiles)
mosaicplot(Travel.vs.Beer, main="", xlab="Travel", ylab="Beer")
```



**Your Turn:** Pick two words that correspond to your interests and compare them using a `mosaicplot`.

**Your Turn:**Describe in words what your `mosaicplot` shows you?

###Analyze statistical significance using `prop.test`. Here we are going to evaluate statistical significance of the difference in the use of words between genders by using a two-sample proportions test. We will first use the word "football" and see if there is statistical significance.

**Your Turn:**Write out a hypothesis test that will test the statistical significance of the word "football" between the two genders.

$H_0$ :

$H_A$ :

```
profiles$Has.Football <- str_detect(essays, "football")
results <- tally(~ Has.Football + sex, data=profiles)
results
```

```
##              sex
## Has.Football     f     m
##         TRUE   741  1298
##         FALSE 23376 34531
```

```
prop.test(x=results[1, ], n=colSums(results), alternative="two.sided")
```

```
## 
##  2-sample test for equality of proportions with continuity
##  correction
## 
## data:  results[1, ] out of colSums(results)
## X-squared = 13.115, df = 1, p-value = 0.0002929
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.008450382 -0.002554462
## sample estimates:
##     prop 1     prop 2 
## 0.03072521 0.03622764
```

**Your Turn:**What are the results of your test?

**Your Turn:**Now choose another word that relates to an interest of yours, perhaps a word that you used earlier, and create a test to evaluate its statistical significance between gender.

$H_0$ :

$H_A$ :

**Your Turn:**What are the results of your test?

###Most Used Words

Here we will look at commonly used words among each gender.

We must first subset the essays into male and female. Then sort them by frequency used in descending order.

Subset male words.

```
male.words <- subset(essays, profiles$sex == "m") %>%   #create a subset of only male essays
str_split(" ") %>%    #split up essays into individual words seperated by a space
unlist() %>%    #create a vector of words
table() %>%     #compute a frequency table
sort(decreasing=TRUE) %>%   #sort in decreasing order
names()         #display words starting with highest frequency ones first
```

**Your Turn:**Subset female words.(Follow the same variable naming convention established for males)

**Your Turn:**Display the Top 25 most frequently used words for each gender. (Be sure to remove the eval=FALSE when you are ready to evaluate this code chunk)

```
male.words[_:__]      #display top 25 male words

                      #display top 25 female words
```

**Your Turn:**What do you notice about the words? Are these words interesting or would you like to see other more descriptive words?

**Your Turn:**Use the `setdiff` command to view the most frequently used words that were in the male top 500 words, but were not in the female top 500 words.(You may need to use the `?setdiff()` help file if you are unfamiliar with it and its arguments.)

```
setdiff()
```

**Your Turn:**Use the `setdiff` command to view the most frequently used words that were in the female top 500 words, but were not in the male top 500 words.(You may need to use the `?setdiff()` help file if you are unfamiliar with it and its arguments.)

```
setdiff()
```

**Your Turn:**Are there any noticable differences in the words that are used by the different genders? Any other thoughts pertaining to the word lists can be entered in the blank text box below.