Project 10 Solutions

(Abhimanyu Agarwal)

Collaborators: N/A

TA help:

1) Melissa : Helped me go through Question 4 and 5.

Online resources used: N/A

Question 1

```
users <- read.csv("/class/datamine/data/okcupid/filtered/users.csv")
questions <- read.csv("/class/datamine/data/okcupid/filtered/questions.csv", sep=";")

#head(questions)
#The questions dataframe stores data regarding the various questions and available options.
#head(users)
#The user dataframe has just captures various answers.
```

Question 2

```
grep("google", questions$text, ignore.case = T)
```

```
[1] 2672
```

```
questions$text[grep("google", questions$text, ignore.case = T)]
```

```
[1] "Do you Google someone before a first date?"
```

```
# Question is : "Do you Google someone before a first date?"
```

Question 3

```
questions[grep("google", questions$text, ignore.case = T),]
```

```
          X                                        text
2672 q170849 Do you Google someone before a first date?
                    option_1                   option_2 option_3 option_4
2672 Yes. Knowledge is power! No. Why spoil the mystery?
         N Type Order    Keywords
2672 39621    0       descriptive
```

```
#Table
prop.table(table(users$q170849, useNA = "always"))
```

```
No. Why spoil the mystery?   Yes. Knowledge is power!
                 0.3774115                  0.2020886
```

```
                    <NA>
                0.4204999
```

```
#Ratio is ~37% No's and 20% acceptances

#Gives percentage based on Man or Woman
tapply(users$q170849, users$gender2, function(x) {prop.table(table(users$q170849, useNA = "always"))})
```

```
$Man

No. Why spoil the mystery?   Yes. Knowledge is power!
                0.3774115                 0.2020886
                     <NA>
                0.4204999

$Woman

No. Why spoil the mystery?   Yes. Knowledge is power!
                0.3774115                 0.2020886
                     <NA>
                0.4204999
```

```
#The ratio is same for each gender. They are ~37% No's and only ~20% yes

#Gives percetange based on Gender Orientation
tapply(users$q170849, users$gender_orientation, function(x) {prop.table(table(users$q170849, useNA = "a
```

```
$Bisexual_female

No. Why spoil the mystery?   Yes. Knowledge is power!
                0.3774115                 0.2020886
                     <NA>
                0.4204999

$Bisexual_male

No. Why spoil the mystery?   Yes. Knowledge is power!
                0.3774115                 0.2020886
                     <NA>
                0.4204999

$Gay_female

No. Why spoil the mystery?   Yes. Knowledge is power!
                0.3774115                 0.2020886
                     <NA>
                0.4204999

$Gay_male

No. Why spoil the mystery?   Yes. Knowledge is power!
                0.3774115                 0.2020886
                     <NA>
                0.4204999
```

```
$Hetero_female

No. Why spoil the mystery?   Yes. Knowledge is power!
                 0.3774115                   0.2020886
                      <NA>
                 0.4204999

$Hetero_male

No. Why spoil the mystery?   Yes. Knowledge is power!
                 0.3774115                   0.2020886
                      <NA>
                 0.4204999
```

```
#The ratio is same for each gender orientation. They are ~37% No's and only ~20% yes
```

Question 4

```
count_words <- function(my_text) {
    my_split_text <- unlist(strsplit(my_text, " "))

    return(length(my_split_text[my_split_text!=""]))
}

#Making the new column question_length which is put into questions
questions$question_length <- sapply(questions$text, count_words)
```

###Question 5

```
number_of_options <- function(myDF)
{
  table(apply(as.matrix(myDF[ ,3:6]), 1, function(x) {sum(!(x==""))}))
}

#Gives options based on the keyword list
keywordlist <- split(questions, questions$Keywords)
sapply(keywordlist, number_of_options)
```

```
[[1]]

  0   4
590   4

$cognitive

 2  3  4
 7  2 15

$descriptive

  2   3   4
387 191 328

$opinion
```

```
 2  3  4
99 44 46

$politics

  2   3   4
119  76  68

$preference

  2   3   4
262 179 239

$`religion/superstition`

 2  3  4
62 27 46
```

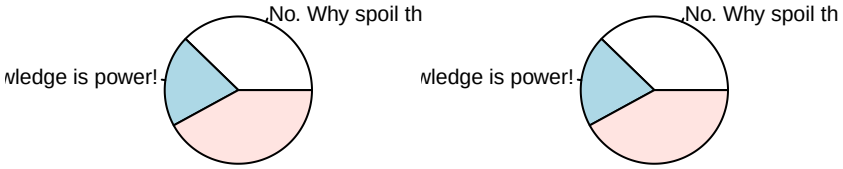
###Question 6

```
par(mfrow=c(2,3))
sapply(tapply(users$q170849, users$gender2, function(x) {prop.table(table(users$q170849, useNA = "always
```

```
$Man
NULL

$Woman
NULL
```

```
#Was just curious to see how the plot comes to be. Essentially i think visual representations are extre
#as it gives a sense of proportions effectively
```



Submitting deliverable: project10.RMD, project10.R and project10.pdf

Pledge

By submitting this work I hereby pledge that this is my own, personal work. I've acknowledged in the designated place at the top of this file all sources that I used to complete said work, including but not limited to: online resources, books, and electronic communications. I've noted all collaboration with fellow students and/or TA's. I did not copy or plagiarize another's work.

> As a Boilermaker pursuing academic excellence, I pledge to be honest and true in all that I do. Accountable together - We are Purdue.