R Practice - dplyr and OCSLS

See if you can answer a few more questions about the Online College Social Life Survey (OCSLS)

Accessing the data

Make sure you load the data by loading the ocsls library

```
library(ocsls)
```

If you get the error "Error in library(ocsls) : there is no package called 'ocsls'", then make sure to install the package with

```
devtools::install_github("mrflick/ocsls")
```

And then try to load the library again.

And also besure to load tidyverse which includes dplyr

```
library(tidyverse)
```

While we will just use the "dates" table, there are many variables in that table. You can start exploring them by browsing the code book available at

http://www.nyu.edu/projects/england/ocsls/codebook/Demographics_index.html

Some notes

- Use the "dates" table unless specified
- Use survey.imputed.year to find the year of the survey

Use dplyr to answer these questions

Note that these can all be answered in one dplyr chain.

1) Which school had the most responses

2) How many schools submitted surveys in 2009?

3) How many people had their most recent date with someone they first met at a summer program?

4) How many respondents reported being in a sorority or fraternity (use "greek")

5) What's the ratio of female to male respondents (use "bio.sex")

6) Which state had the smallest proportion of female respondents (use "state.graduated.high.school" and "bio.sex")

7) Use this function to create a data.frame in R that has the land area in square miles for each of the 50 states

```
states <- rownames_to_column(data.frame(state.x77), "state")
```

Which state at the most number of respondents per square mile (use "state.graduated.high.school" from "dates" and "Area" from the "states" data.frame we just created)

8) At which school is the boy most likely to pay on a date (use "which.sex.paid")

Expected Values

1) U Mass with 3607
2) 12
3) 8
4) 2857
5) 2.208685 females to males
6) Oklahoma with only 45%
7) Massachusetts with 1 per 0.49 sq miles
8) Arizona with 74.8%

Possible Answers

There may be multiple ways to do things in R with dplyr and there are different ways to interpret the question, so these are not necessarily the "correct" answers; they are just possible answers.

1) Which school had the most responses

```
dates %>% count(school) %>% arrange(-n) %>% top_n(1)
# 1 U Mass  3607
```

2) How many schools submitted surveys in 2009?

```
dates %>% filter(survey.imputed.year==2009) %>% count(school) %>%
nrow()
# 12
```

3) How many people had their most recent date with someone they first met at a summer program?

```
dates %>% filter(tolower(first.met.at.specify)=="summer program")
%>% nrow()
# [1] 8
```

4) How many respondents reported being in a sorority or fraternity (use "greek")

```
dates %>% summarize(frat=sum(greek, na.rm=TRUE))
1  2857
```

5) What's the ratio of female to male respondents (use "bio.sex")

```
dates %>% summarize(males=sum(bio.sex=="Male", na.rm=T),
females=sum(bio.sex=="Female", na.rm=T), ratio=females/males)
#  males females    ratio
#1  7461   16479 2.208685
```

6) Which state had the smallest proportion of female respondents (use "state.graduated.high.school" and "bio.sex")

```
dates %>% select(state=state.graduated.high.school, bio.sex) %>%
na.omit() %>% group_by(state) %>%
summarize(males=sum(bio.sex=="Male"),
females=sum(bio.sex=="Female"), prop_f=females/(males+females))
%>% top_n(-1)
# Oklahoma    11       9   0.45
```

7) Which state at the most number of respondents per square mile (use "state.graduated.high.school" from "dates" and "Area" from the "states" data.frame we just created)

```
dates %>% select(state=state.graduated.high.school) %>% na.omit()
%>% mutate(state=tolower(state)) %>%count(state) %>%
inner_join(states %>% mutate(state=tolower(state))) %>%
mutate(per_mile=n/Area) %>% top_n(1, per_mile) %>%
select(state_permile)
#          state  per_mile
# 1 massachusetts 0.4904166
```

8) At which school is the boy most likely to pay on a date (use "which.sex.paid")

```
dates %>% select(school, which.sex.paid) %>% na.omit %>%
group_by(school) %>% summarize(n=n(),
perc_boy=mean(which.sex.paid=="Boy paid", na.rm=TRUE)) %>%
top_n(1)
#    school     n  perc_boy
# 1 Arizona  1044 0.7480843
```