Homework 3: Stats with R

Mike Lopez

September 2017

General instructions for homeworks:

- Make a new R Markdown file (.Rmd) referring to the assignment on the course Github page
- Change the heading to include your author name
- Save the R Markdown file (named as: [MikeID]-[Homework01].Rmd – e.g. "mlopez-Lab01.Rmd") to somewhere where you'll be able to access it later (zip drive, My Documents, Dropbox, etc)
- Your file should contain the code/commands to answer each question in its own code block, which will also produce plots that will be automatically embedded in the output file
- **Each answer must be supported by written statements (unless otherwise specified) as well as any code used**: In other words, if the answer is `24`, you should write "The answer is 24" (as opposed to just showing the code and output).
- Include the names of anyone you collaborated with at the top of the assignment
- I recommend copying the raw .Rmd code from the Github page as a start
- Homeworks are due at the start of class – please print the HTML and hand in.

The data set that we'll be working with for this assignment is found at:

```
library(okcupiddata)
names(profiles)
?profiles
```

Refer to the help screen for information regarding each variable.

Part I: commands

1. Filter all profiles that are
    i. at least 72 inches tall
    ii. have a gemini astrological sign.
    iii. smoke sometimes
    iv. smoke sometimes and are at least 72 inches tall
    v. home a gemini or aquarius astrological sign
    vi. are between 72 and 76 inches tall, inclusive

For this question, you do not need to show the data frame – just the code used to identiy these profiles.

2. Using a data set only with gemini's who smoke sometimes, identify height of the shortest person and the tallest person.

3. Create the following new variables
    i. `hard.pass`, which represents any profile who responds `dislikes dogs and dislikes cats` to the `pets` variable
    ii. `good.luck`, which represents any profile who responds `has dogs and has cats` to the `pets` variable
    iii. `inches.difference`, which represents the difference in height between your professor (77 inches) and the profile of interest (note: use your height here, in inches)

that we'll be using for this lab are (i) `mpg` and (ii) a random sample of `flights`, the latter of which can be found in the `nycflights13` package. See class notes for exact details on the flights data set, or enter `?flights` to use the Help tab.

4. Calculate the average, median, and standard deviation of income among all of the profiles. Note – there are people with missing incomes, in which case adding `na.rm = TRUE` to your code can ensure that you drop subjects whose incomes are missing.

5. Calculate the average income based on each type of pet preference, and arrange in order to find the pet category with the largest and smallest income levels. Is there any pattern?

6. Which drink type tends to have the oldest profiles? Use techniques discussed above to create a new data set, and create a visualization that justifies your findings.

7. Find the job type that has the largest percent of profiles that drink `desparately`.