COMPSCIX 415.2 Homework 5/Midterm

Robert Clements

DUE DATE: Mar 6, 2018 @ 6:15PM

In this assignment/midterm...

We will review all of these topics:

- RStudio and R Markdown/Notebooks
- The tidyverse suite of packages
- R basics
- Data import/export
- Visualization
- Data munging and wrangling
- EDA
- git and Github (clone, commit, push)

Remember to work on this independently. If you have questions you can reach me on Canvas or through email at robert_clements@berkeley.edu.

Remember to make your document look good, which means you may need to change some settings on the figure sizes and locations, use markdown syntax to create headings or to format your text (use the cheatsheet), and you may want to play with the different themes.

Use complete sentences, and **divide your work in a logical way**. Remember that the whole point of doing reproducible analysis in R Markdown is so that a complete stranger can take your results, *understand them*, and reproduce them.

Remember to save and knit often. Commit when you've completed a big chunk of work or when you are done for the day and will be resuming later.

What to Turn In

For this assignment you have two choices:

You can upload a pdf document (you will have to install latex);
You can upload a standalone html file.

To complete this midterm you will need...

Access to the internet.
The `tidyverse` package installed.
RStudio and git/Github.

To start your assignment

1. Go to File -> Recent Projects

2. Click on the Project that you created during Homework 1. This project should be the one that is already under git version control.

3. RStudio will switch to that project and the Git pane should appear.

4. Go to File -> New File and choose **R Markdown**.

5. Change the title ("COMPSCIX 415.2 Homework 5/Midterm") and add your name and the date to the YAML header.

6. Save the file in the same folder (or create a subfolder) with your other HW assignments and give it the name **firstname_lastname_midterm.Rmd**.

7. Knit your document into an html or pdf document.
8. Go to the Git pane and commit both your Rmd and html (or pdf) files by clicking on the checkboxes next to the file names and hitting the Commit button. Write a useful message, and hit the commit button.

Exercises (Total Points - 30)

RStudio and R Markdown (3 points)

1. Use markdown headers in your document to clearly separate each midterm question and add a table of contents to your document.

The tidyverse packages (3 points)

By now you've used at least five different packages from the tidyverse for plotting, data munging, reshaping data, importing/exporting data, and using tibbles (the `tibble` package is used for this without you even realizing it's there).

1. Can you name which package is associated with each task below?

Plotting -
Data munging/wrangling -
Reshaping (speading and gathering) data -
Importing/exporting data -

2. Now can you name two functions that you've used **from each package** that you listed above for these tasks?

Plotting -
Data munging/wrangling -
Reshaping data -
Importing/exporting data (note that `readRDS` and `saveRDS` are base R functions) -

R Basics (1.5 points)

1. Fix this code *with the fewest number of changes possible* so it works:

```
My_data.name___is.tooOOooLong! <- c( 1 , 2    , 3 )
```

2. Fix this code so it works:

```
my_string <- C('has', 'an', 'error', 'in', 'it)
```

3. Look at the code below and comment on what happened to the values in the vector.

```
my_vector <- c(1, 2, '3', '4', 5)
my_vector
```

```
## [1] "1" "2" "3" "4" "5"
```

Data import/export (3 points)

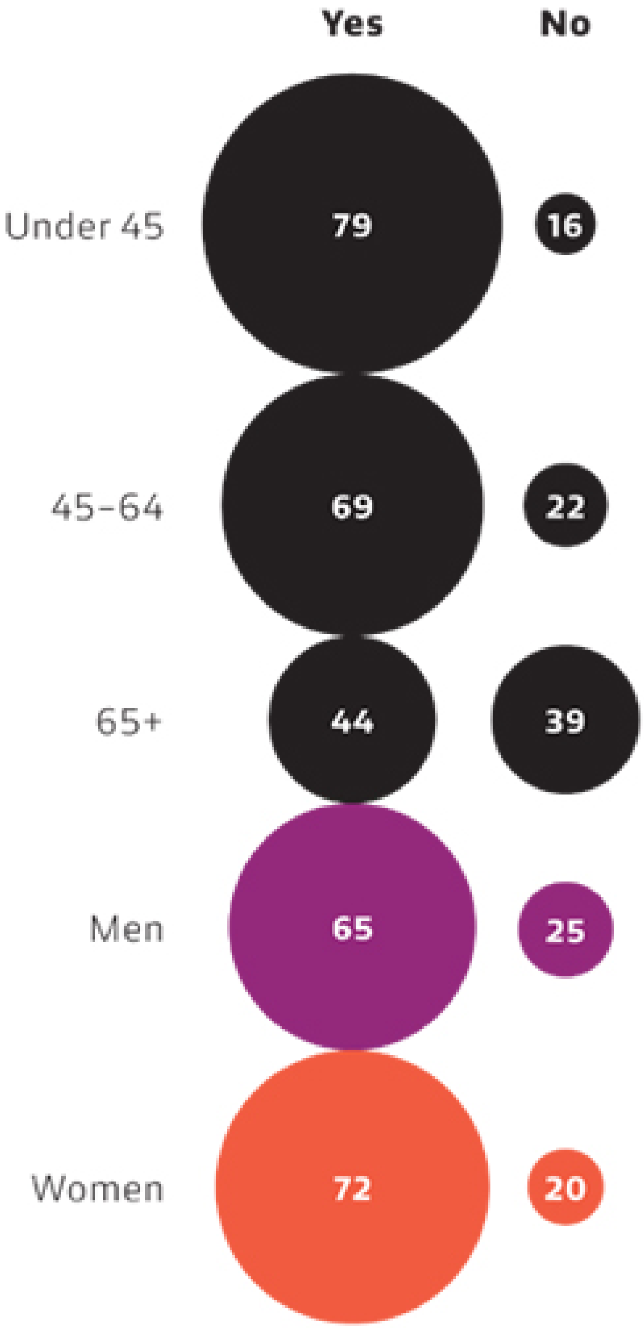
1. Download the rail_trail.txt file from Canvas (in the Midterm Exam section here) and successfully import it into R. Prove that it was imported successfully by including your import code and taking a `glimpse` of the result.
2. Export the file into an R-specific format and name it "rail_trail.rds". Make sure you define the `path` correctly so that you know where it gets saved. Then reload the file. Include your export and import code and take another `glimpse`.

Visualization (6 points)

1. Critique this graphic: give **only three** examples of what is wrong with this graphic. Be concise.

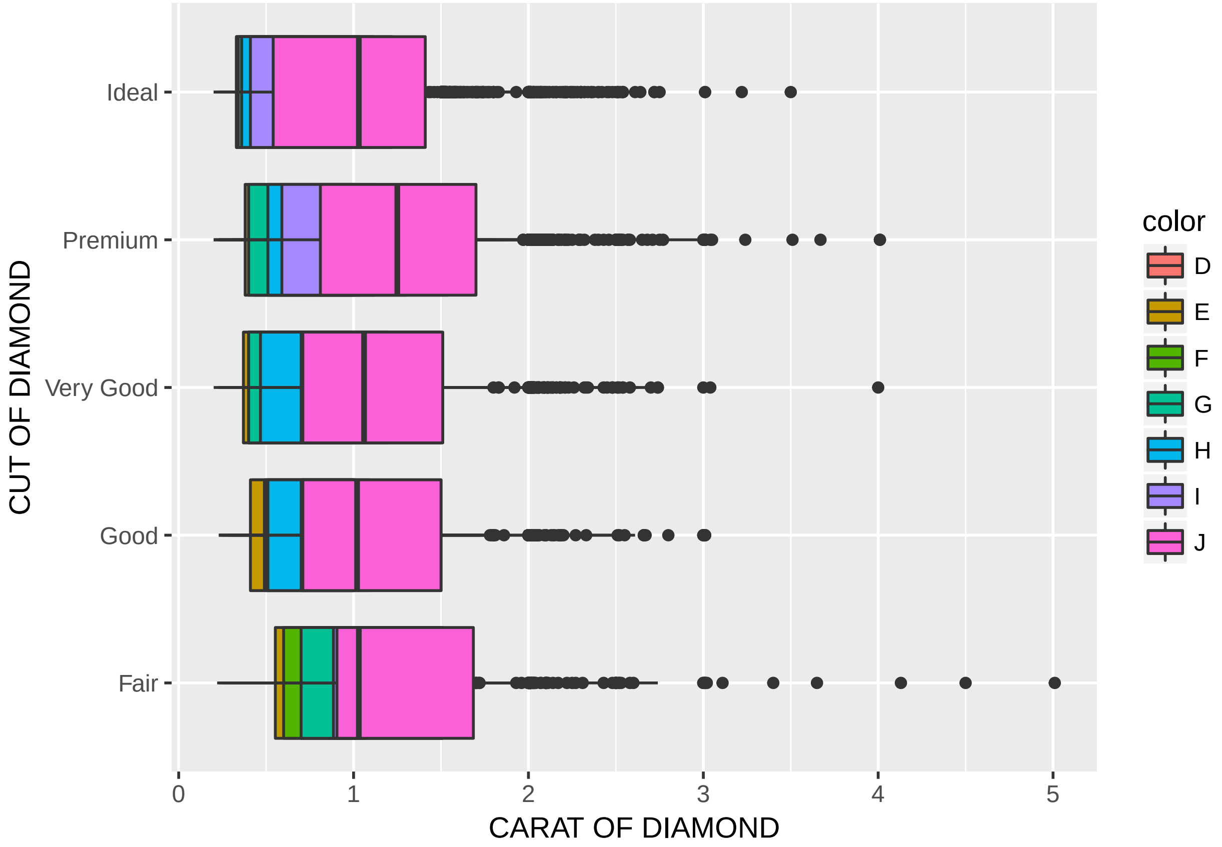
MRS. PRESIDENT

Percentage of respondents who say it is likely that a woman will be president in their lifetime.



Source: CBS News, June 2008

2. Reproduce this graphic using the `diamonds` data set.



3. The previous graphic is not very useful. We can make it much more useful by changing one thing about it. Make the change and plot it again.

Data munging and wrangling (6 points)

1. Is this data "tidy"? If yes, leave it alone and go to the next problem. If no, make it tidy. *Note: this data set is called* `table2` *and is available in the tidyverse package. It should be ready for you to use after you've loaded the tidyverse package.*

```
table2
```

```
## # A tibble: 12 x 4
##    country      year type            count
##    <chr>       <int> <chr>           <int>
##  1 Afghanistan  1999 cases             745
##  2 Afghanistan  1999 population   19987071
##  3 Afghanistan  2000 cases            2666
##  4 Afghanistan  2000 population   20595360
##  5 Brazil       1999 cases           37737
##  6 Brazil       1999 population  172006362
##  7 Brazil       2000 cases           80488
##  8 Brazil       2000 population  174504898
##  9 China        1999 cases          212258
## 10 China        1999 population 1272915272
## 11 China        2000 cases          213766
## 12 China        2000 population 1280428583
```

2. Create a new column in the `diamonds` data set called `price_per_carat` that shows the price of each diamond per carat (hint: divide). Only show me the code, not the output.

3. For each `cut` of diamond in the `diamonds` data set, how many diamonds, and what proportion, have a price > 10000 and a carat < 1.5? There are several ways to get to an answer, but your solution **must** use the data wrangling verbs from the tidyverse in order to get credit.

    - Do the results make sense? Why?
    - Do we need to be wary of any of these numbers? Why?

EDA (6 points)

Take a look at the `txhousing` data set that is included with the `ggplot2` package and answer these questions:

1. During what time period is this data from?
2. How many cities are represented?
3. Which city, month and year had the highest number of sales?
4. What kind of relationship do you think exists between the number of listings and the number of sales? Check your assumption and show your work.
5. What proportion of `sales` is missing for each city?
6. Looking at only the cities and months with greater than 500 sales:
    - Are the distributions of the median sales price (column name `median`), when grouped by city, different? The same? Show your work.
    - Any cities that stand out that you'd want to investigate further?
    - Why might we want to filter out all cities and months with sales less than 500?

Git and Github (1.5 points)

To demonstrate your use of git and Github, at the top of your document put a hyperlink to your Github repository.

Once you are finished with your midterm, commit your final changes with the comment "finished the midterm - woohoo" and push your R Markdown file and your html or pdf file to Github.

Turn in your completed midterm

This week you should turn in your midterm by uploading it to Canvas.