Homework for Intro Module

Due Tuesday, June 28, 2022 at 10am

- You will find the datasets to do this homework in the directory https://web.stanford.edu/class/stats101/data/.
- The assignment is due on `gradescope`. Please upload the `PDF` that you obtain by knitting the `Rmd` file that contains your `R` code and your text answering other questions. So this uploaded file will also show any output that `R` produces in addition to your code. Learning to create documents this way is part of this assignment. When uploading your document, please make sure to assign each page to the correct question. To create a new RMarkdown file you can follow these steps in RStudio: Click "File" in the bar at the top of RStudio, then click "New File" and choose "R Markdown". Click on "PDF" as default output format. You can then save this newly created file in your folder. To "knit" the document, locate the "Knit" button in RStudio and the small triangle to the right of it and choose "Knit to PDF". Note that in order to successfully knit to PDF you might need to install a LaTeX interpreter, e.g. MikTeX. After the PDF has been knitted successfully, a pdf file will be created in the folder where you saved your RMarkdown file. This is the pdf you want to submit (not the "preview" RStudio opens after successful knitting).
- Getting started with R, RMarkdown might take some time. Especially problem 3 of this homework might take some time. Please get started early on this homework. If you get stuck, we are happy to help you in office hours or on Ed.

1. Some Calculations

This question asks you to have `R` do a few calculations. (*Hint*: all of these can be done without writing a loop.)

(a) Create a vector called `x` with 10 different numbers, some positive, some negative. Create another vector `y` the same way but with different numbers. (*Hint*: you can use `c()` to instantiate a vector)

(b) What is the mean of `x`? What is the mean of `y`?

(c) Create two vectors, one that stores the difference between `x` and its mean and another that does the same for `y`. Call these `xd` and `yd`, respectively.

2. Reproducible research

Using `Rmd` to record and annotate your research results is taking one step to assure that your research is reproducible. If you search the web for this term, you will see that this is a central focus of contemporary science. To have an introduction to the topic, read the article An invitation to reproducible computational research. Summarize three key ideas discussed in it. A couple of sentences is sufficient.

3. Sprint data

The link to the datasets at the top of this homework contains the datasets **100men** and **100women** with the history – referred to in track and field lingo as the "progression" – of world records in the 100 meters for men and women. Download them into a folder `data` on your computer. Then run the code below. If you get error messages that the package "lubridate" or "ggplot" is not found etc, try using the command "install.packages()" (inside the bracket put the package you would like to install) and then try using the "library()" command again after installing.

```
library('lubridate')
```

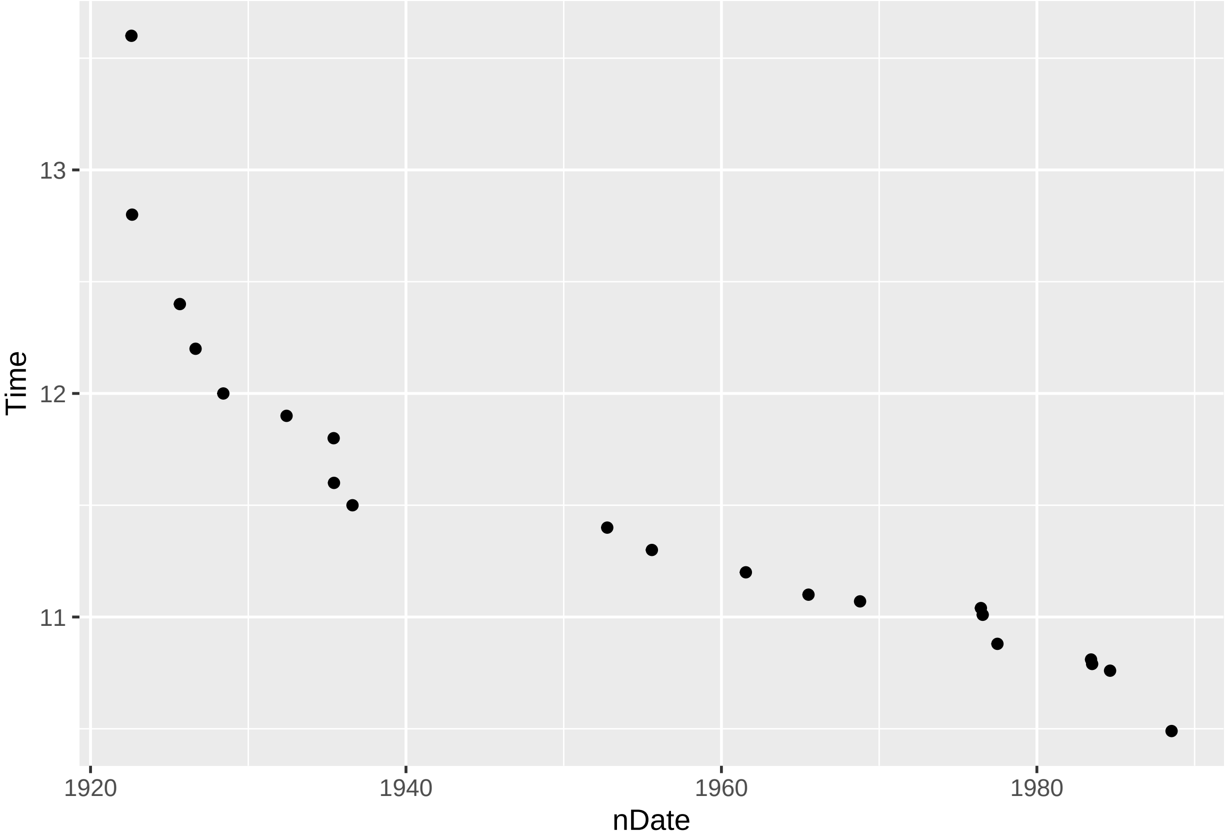
```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
men100 <- read.csv("data/100men",sep='\t')
women100 <- read.csv("data/100women",sep='\t')
women100$nDate <- as_date(as.character(women100$Date),format='%b %d, %Y')
require('ggplot2')
```

```
## Loading required package: ggplot2
```

```
ggplot(women100,aes(x=nDate,y=Time)) + geom_point()
```



(a) Read in the datasets and extract the nationalities for each sprinter. This means that you should e.g. create a vector holding the nationalities for men and women. Note that the spellings of different nations might differ between the men and the women dataset. You might need to do some "data cleaning" such that the nation names in both datasets are equal to prepare you for the next question.

(b) Create a graphical display of the nationalities of the different sprinters, distinguishing men and women. Hint: the "table()" function might be helpful to get the frequencies of each nation for the women and the men. After your data cleaning (e.g. nations that appear in both women and men dataframes have the same spelling, you might want to use the “merge()” function to combine the frequency tables for the men and the women nationalities).

(Note: this **does not** mean you need to make a map, you can use a table or bar char. Ideally you should not generate two separate displays (one for men, one for women), but **one display for both**. However, if following the hint above gives you too much trouble, we will accept two (correct) separate plots or tables (one for men and one for women) for full credit as well. Nevertheless, please try to create one display for both.)

Describe any patterns that you see.

(c) Are males or females improving more quickly over time? Are women “catching up” to men? Make some plot(s) to examine this question. Note: This is an open-ended question and you can use any type of plot that you think sheds light on the question.

(d) The file **200men** contains records for the 200m sprint by men. What strikes you as different/similar in comparing the 200m and the 100m sprints? Support your analysis with a plot or table. (Note: Again, this is open ended and you can use any type of display that you think sheds light on the question).

4. “Big Data Religion”

Read the article from Gil Press (note that he has written many pieces on the subject and you might enjoy reading more than one). Write a short discussion of the article that emphasizes the conceptual points - a couple of sentences is sufficient. Make sure to include in your summary two following points:

(a) Why does the media associate Big Data with a “religion” rather than “science”?
(b) What elements of Snow’s investigation on the mode of transmission of Cholera differ from the mode of research described as “big data religion”?