DCF Week 6 Warm-Up & Assignment

Data and Computing Fundamentals

Create your Rmd file from the DCF assignment template, saving it as `Week-6-Warmup-XXX.Rmd`.

Is death seasonal?

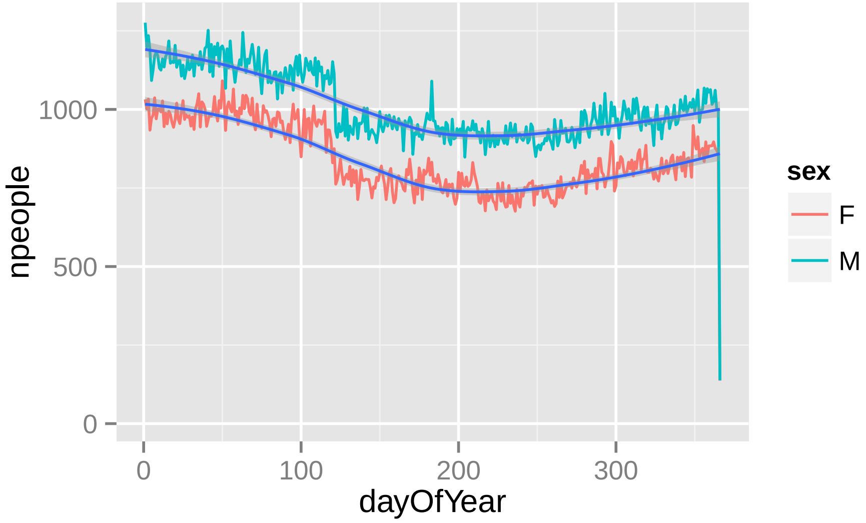
You're going to do some data cleaning. The data originally come from https://health.data.ny.gov/Health/Genealogical-Research-Death-Index-Beginning-1957/vafa-pf2s, where they are available as a CSV file.

The data from the site were downloaded, creating a CSV file of size 47.4 MB. In order to make cleaning the data more convenient, the CSV file was pre-processed to include just five of ten original variables. This was saved in Rda format, ending up at 2.9 MB. That's the file you will be using. Load it with this command:

```
load( url( "http://tinyurl.com/m4o4n2b/DCF/Deaths.rda" ) )
```

Take a look at the data with `str()`. Note that it consists of five variables with lengthy names. Four of the five variables are character strings.

You're going to clean these data to make it possible to graph them like this:



Some curious observations:

- What's going on at about day 120? Why the sudden drop in the number of deaths?
- What's going on at the very end of the year?

You don't have to fix these problems (yet), just explain what they are.

Make the names convenient

The names given in the original file are very long. It's helpful to shorten them up, like this:

```
Deaths <- Deaths %>% head %>%
  select( name=`Decedent First Name`,
          age=`Decedent Age`,
          ... and so on for the other three vars.)
```

Note that the original variable names are quoted with a backquote (‘). This is for technical reasons, but you must do it when variable names contain spaces or other punctuation characters.[1]

How many sexes?

Check the possible values for sex. Why are there so many? Fix them so that the different character-string versions of the same sex are unified.

Hint: To delete characters, even blank characters, you can use `gsub()`.

Names of People

Look at a few of the first names of the people. The character strings holding the names have several blank spaces after the name. These are called *trailing spaces*.

Get rid of the trailing spaces, while retaining any spaces within the names, as in Lu Ann.

While you're at it, get rid of leading spaces (if any), the blanks at the front of the string.

Age is not always in years

The "units" variable shows the units with which the age has been recorded. How many distinct levels of units are there? Unify any that have trivial typographical differences.

Day of Week, Day of Year

The date of death is a character string in the original data. Turn this into a date object using the appropriate one of the relevant functions from the `lubridate` package: `ymd()`, `mdy()`, `dmy()`.

You can also find the day of the year (January 1 is day 1, February 1 is day 32, and so on) and the day of the week (Sunday, Monday, . . .) using the `yday()` and `wday()` functions.

Make your plot

Make the plot shown at the top of this section.

The sharp drop near day 120 is an artifact. Figure out what causes it and fix it.

[1] R lets you use just about anything as a variable name, so long as the name is enclosed in backquotes.