Homework #3

For questions 1-4 in this problem set, we will work with a dataset on dogs of New York City, found here: https://project.wnyc.org/dogs-of-nyc/

Please use the "NYCdogs.csv" version found in Files/Data folder on CourseWorks, which includes a Group column. If you already did some of the questions that didn't require the Group column, you do not have to redo them.

Background: The dataset is dated June 26, 2012. Although the data were originally produced by the NYC Department of Mental Health and Hygiene, it no longer seems to be available on any official NYC web site. (There is a 2016 dataset on dog licenses with different variables available here: https://data.cityofnewyork.us/Health/NYC-Dog-Licensing-Dataset/nu7n-tubp). Also of note is the fact that this dataset has 81,542 observations. The same summer, the New York City Economic Development Corporation estimated that there were 600,000 dogs in New York City (source: https://blog.nycpooch.com/2012/08/28/how-many-dogs-live-in-new-york-city/) Quite a difference! How many dogs were there really in 2012?!? Might be an interesting question to pursue for a final project, but for now we'll work with what we've got.

1. Missing Data

(a) Create a bar chart showing percent missing by variable.

(b) Use the `extracat::visna()` to graph missing patterns. Interpret the graph.

(c) Do `dog_name` missing patterns appear to be associated with the *value* of `gender`, `Group` *or* `borough`?

2. Dates

(a) Convert the `birth` column of the NYC dogs dataset to `Date` class (use "01" for the day since it's not provided). Create a frequency histogram of birthdates with a one-month binwidth. (Hint: don't forget about base R.) What do you observe? Provide a reasonable hypothesis for the prominent pattern in the graph.

(b) Redraw the frequency histogram with impossible values removed and a more reasonable binwidth.

3. Mosaic plots

(a) Create a mosaic plot to see if `dominant_color` depends on `Group`. Use only the top 5 dominant colors; group the rest into an "OTHER" category. The last split should be the dependent variable and it should be horizontal. Sort each variable by frequency, with the exception of "OTHER", which should be the last category for dominant color. The labeling should be clear enough to identify what's what; it doesn't have to be perfect. Do the variables appear to be associated? Briefly describe.

(b) Redraw with the "OTHER" category filtered out. Do the results change? How should one decide whether it's necessary or not to include an "OTHER" category?

4. Maps

Draw a spatial heat map of the percent spayed or neutered dogs by zip code. What patterns do you notice?

5. Time Series

(a) Use the `tidyquant` package to collect information on four tech stocks of your choosing. Create a multiple line chart of the closing prices of the four stocks on the same graph, showing each stock in a different color.

(b) Transform the data so each stock begins at 100 and replot. Choose a starting date for which you have data on all of the stocks. Do you learn anything new that wasn't visible in (a)?

6. Presentation

Imagine that you have been asked to create a graph from the Dogs of NYC dataset that will be presented to a very important person (or people). The stakes are high.

(a) Who is the audience? (Mayor DeBlasio, a real estate developer, the voters, the City Council, the CEO of Purina...)

(b) What is the main point you hope someone will take away from the graph?

(c) Present the graph, cleaned up to the standards of "presentation style." Pay attention to choice of graph type, if and how the data will be summarized, if and how the data will be subsetted, title, axis labels, axis breaks, axis tick mark labels, color, gridlines, and any other relevant features.