Homework 04: Data frames and data wrangling

Please read the entire chapter on data transformation from
R for Data Science before starting this homework.

This homework relies on the `nycflights13` package, which contains several data frames, including:

- `airlines`
- `airports`
- `flights`
- `planes`
- `weather`

Loading `nycflights13` puts these data frames on the search path.

Setup

Load packages (do this every time)

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.6
v forcats   1.0.1     v stringr   1.6.0
v ggplot2   4.0.1     v tibble    3.3.1
v lubridate 1.9.4     v tidyr     1.3.2
v purrr     1.2.1
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
```

```
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

```
library(nycflights13)
```

Note - if you don't have these installed you will need to first install.packages('nycflights13')

Question 1: Filtering

Make a plot of **air time vs distance** (air time on the y-axis, distance on the x-axis) for all flights that meet the following criteria:

- originate from LaGuardia airport (`"LGA"`)
- departed on the 16th of the month
- have a flight distance of less than 2000 miles

Your code

```
# Filter flights according to the criteria above,
# then make the required plot.
```

Brief written response

In 2–4 sentences, describe what you observe in the plot. (For example: Is the relationship roughly linear? Are there any clear outliers?)

Your answer here:

Question 2: Dealing with NAs

Make a data frame of all rows of `flights` that have values for *both* `arr_time` and `dep_time` (i.e., neither value is `NA`).

Your code

```
# Create a new data frame that removes rows where arr_time or dep_time is NA.
```

Filtering NAs (conceptual)

`ggplot()` will automatically remove `NA` values from a plot, but it emits a warning message about it. You *could* silence warnings using chunk options, but instead:

Brief written response

Explain (in words) how you could prevent those `NA` values from appearing in the plot in the first place.

Your answer here:

Question 3: Adding columns

Create a data frame of **average flight speeds**, based on `air_time` and `distance`.

Then make either:

- a histogram, **or**
- a density plot

If you like, you may break the data out (e.g., by airline or another variable) in a way that you think makes sense.

Your code

```
# Create a new column for average speed.
# (Hint: think carefully about units - air_time is in minutes.)
# Then make a histogram or density plot.
```

Brief written response

Describe the main features of your speed distribution.

Your answer here:

Rendering and submission (GitHub)

Canvas contains the course GitHub link and instructions for **forking** the course repository.

In this file, your job is to:

1. Complete the homework by adding code + written responses above.
2. Render this Quarto file to **HTML or PDF**.
3. Move the rendered file(s) to the /doc folder (note that, if you render to html you will also have to move your hmk_04_data_frames_files folder)
4. Add/commit both:
    - this edited `.qmd` file
    - the rendered output file(s) (HTML or PDF)
5. Push your changes to your fork on GitHub.
6. Open a Pull Request back to the course repository.

What to submit on Canvas

On Canvas, submit a link to your GitHub fork repository (not files).