21FA DATA 202 Midterm 1

In a preparatory exercise you worked with data from the `nycflights13` package, which contains data about flights that departed from New York City (i.e., airports EWR, JFA, and LGA) in 2013. We will be working with the `flights` table, which has one row per flight. Here is a sample containing 10 rows of that table:

```
set.seed(0)
flights_sample <- flights %>%
  slice_sample(n = 10) %>%
  select(year, month, day, hour, carrier, origin, dest, dep_delay, arr_delay)
flights_sample
```

year	month	day	hour	carrier	origin	dest	dep_delay	arr_delay
2013	6	13	19	UA	LGA	IAH	57	154
2013	1	16	7	9E	EWR	CVG	-5	30
2013	9	17	10	EV	JFK	IAD	-6	-18
2013	4	8	16	DL	JFK	DTW	-6	-20
2013	11	26	6	B6	JFK	FLL	-5	4
2013	4	12	14	DL	LGA	DTW	25	13
2013	9	4	18	AA	EWR	LAX	-5	-33
2013	2	2	12	B6	JFK	PBI	34	27
2013	7	31	18	DL	JFK	ATL	15	-1
2013	1	2	6	EV	LGA	MEM	10	19

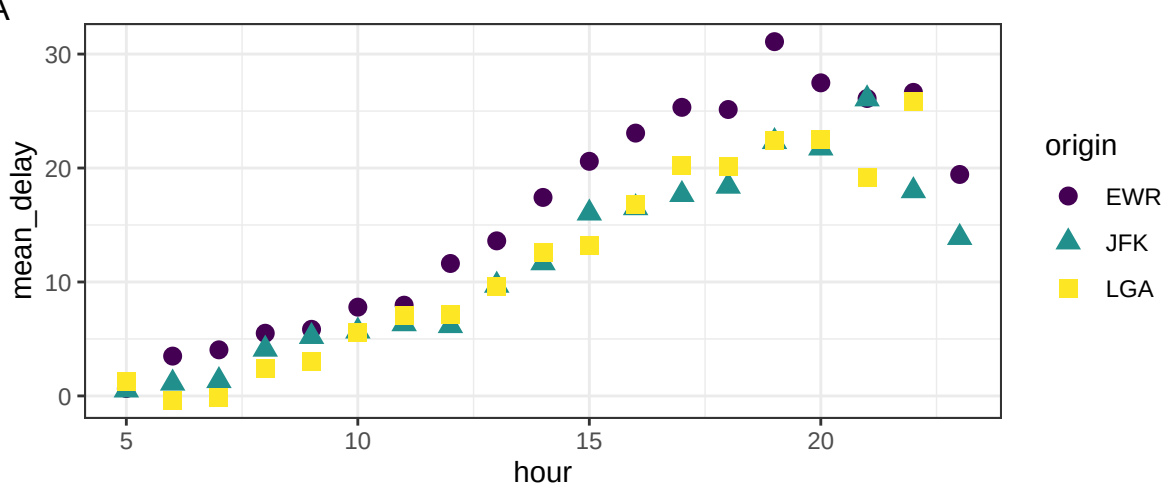
The `nycflights13` package also provides a table of `airlines`:

carrier	name
9E	Endeavor Air Inc.
AA	American Airlines Inc.
AS	Alaska Airlines Inc.
B6	JetBlue Airways
DL	Delta Air Lines Inc.
EV	ExpressJet Airlines Inc.
F9	Frontier Airlines Inc.
FL	AirTran Airways Corporation
HA	Hawaiian Airlines Inc.
MQ	Envoy Air
OO	SkyWest Airlines Inc.
UA	United Air Lines Inc.
US	US Airways Inc.
VX	Virgin America
WN	Southwest Airlines Co.
YV	Mesa Airlines Inc.

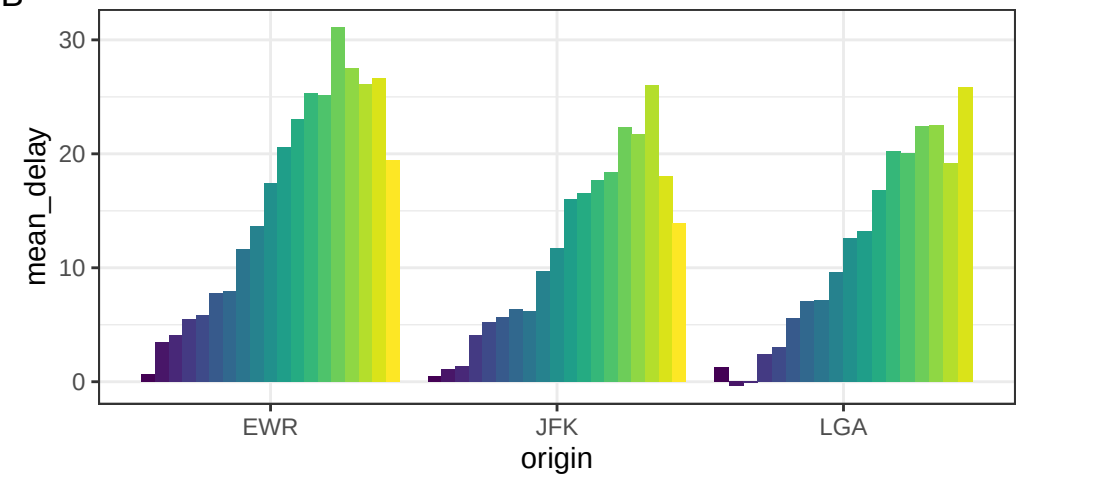
Visualization Exercises

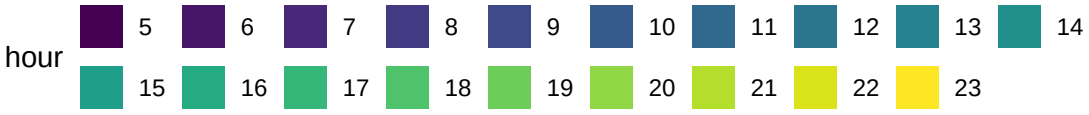
Consider the following two plots. Both are based on the full `flights` data frame, but the data may have been transformed in some way before the plot.

A



B





1. What type of plot is plot A?

2. What type of plot is plot B?

3. Make a table of the aesthetic mappings for plot A.

4. Make a table of the aesthetic mappings for plot B.

5. Suppose plot A was constructed by supplying `tableA` to a `ggplot` command, and that nothing fancy happens in the `ggplot` (no `geom`s that count rows, no `stat_*` commands, etc.).
   a. Give a plausible example of the first 3 rows of `tableA` (i.e., `head(tableA, 3)`). Omit any columns that aren't used in the plot.

   b. How many rows do you expect `tableA` to have? Write (but do not evaluate) the multiplication expression that you would enter into a simple calculator to give the this result.

6. Repeat problem 5 for plot B, with the same instructions (abbreviated below).
   a. Give a plausible example of the first 3 rows of `tableA`.

   b. How many rows do you expect `tableA` to have?

7. Give an example of something that you can notice about the data (i.e., a comparison you can make) *more easily in plot A than in plot B*.
   a. What is the specific observation you make? (e.g., "The mean delay for XXX is higher for . . . ")

   b. What about the design of the visualization makes that comparison easier in plot A?

8. Repeat problem 7, but for plot B instead.
   a. What is the specific observation you make? (e.g., "The mean delay for XXX is higher for . . . ")

   b. What about the design of the visualization makes that comparison easier in plot B?

9. Identify a change to make to plot A that would improve it.
   a. Circle or draw an arrow to the part of the plot that you would change.
   b. Next to the plot, sketch what that part of the plot would look like after your change.
10. Repeat problem 9 for plot B.

Data Transformation Exercises

For each of the following pipelines:

- Specify the *shape*: the total number of rows and columns in the result.
- Give the first two rows of the result.

All of the pipelines use the table `flights_sample` shown on the first page. Note that it only has 10 rows, so you can perform all of these operations by hand.

Each exercise is 2 points for **correct** and **complete** column names, 1 point for correct number of rows, and 1 point for correct content.

Exercise 11 Shape:

```
flights_sample %>%
  arrange(dep_delay)
```

Exercise 12 Shape:

```
flights_sample %>%
  filter(hour > 16 & arr_delay > 0)
```

Exercise 13 Shape:

```
flights_sample %>%
  group_by(origin) %>%
  summarize(count = n()) %>%
  arrange(desc(count))
```

Exercise 14 Shape:

```
flights_sample %>%
  filter(origin == "LGA") %>%
  left_join(airlines, by = "carrier")
```

Exercise 15 Shape:

```
flights_sample %>%
  pivot_longer(
    cols = c('origin', 'dest'),
    names_to = "role",
    values_to = "airport") %>%
  group_by(role, airport) %>%
  summarize(count = n())
```