R Practice - dplyr and NYC Flights

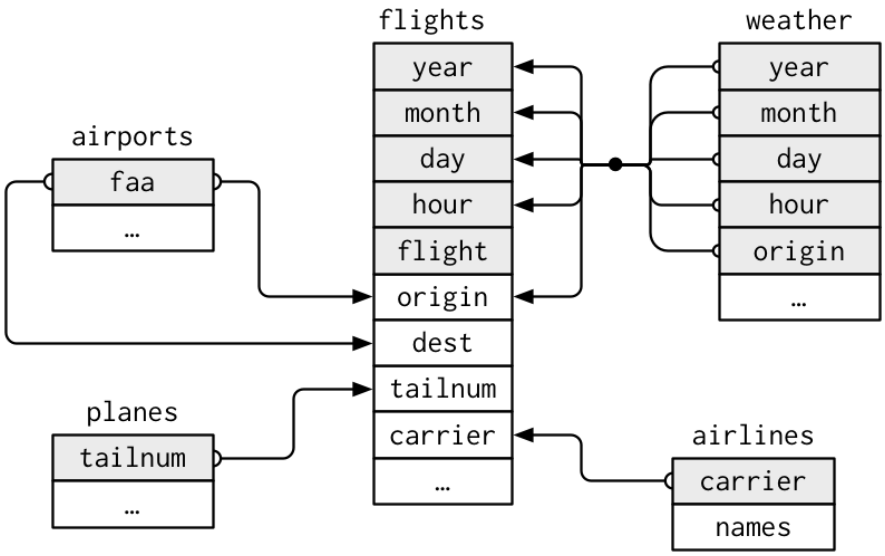
See if you can answer a few more questions about the NYC flights data set

Accessing the data

Make sure you load the tidyverse package which includes dplyr as well as the nycflights13 package to get the flights data

```
library(tidyverse)
library(nycflights13)
```

Recall that the library includes the following tables



Use dplyr to answer these questions

Note that these can all be answered in one dplyr chain.

1) What destination received most flights in June?

2) Which carrier had the greatest average distance per flight?

3) Which flight traveled the fastest (overall miles per hour)?

4) What day had the largest average arrival delay for all flights? (only count positive values)

5) What was the total distance for all flights in January? What was the average distance per flight?

6) What day of the week saw the most flights? (Hint, use the "time_hour" column which is a date/time column and use google to find a function to give you the weekday for a date).

7) What was the average number of seats and engines on the plains that left from NYC on July 4? (The "seats" and "engines" data comes from the the planes table; Warning: Be careful with your joins!)

8) How many airlines do not have the word "air" somewhere in their name? (Hint, try gooling for a function that can do string matching in R)

9) What was the most common plane model to fly out of NYC in October (there is a "model" column in the planes table)?

10) How many planes (tailnum) only flew one route (flight) but flew that route more than 10 times?

11) Which scheduled departure hour (use the "hour" column) had the largest proportion of flights delayed (dep_delay) longer than 5 min?

12) Which flight(s) had the greatest scheduled length (time between scheduled departure and arrival)? (Warning: Take notice how the values of `sched_arr_time and sched_dep_time` are formatted in the table).

Expected Values

1) Chicago O'Hare (ORD) with 1547
2) Hawaiian Airlines Inc. with an average of 4983 miles per flight
3) Flight 1499 to ATL flew 762 miles with average speed of 703.3846
4) July 10, with an average delay of 110 minutes
5) 27,188,805 total miles with an average of 1006.844 miles per flight
6) Monday with 50690
7) Seats: 140.6581 Engines: 1.991974
8) Only Virgin America
9) The A320-232 with 3717 flights
10) 12 planes only flew one route but did so more than 10 times
11) 9PM with 46.3%
12) Flight 51 with 6 hours 40 minutes

Possible Answers

There may be multiple ways to do things in R with dplyr and there are different ways to interpret the question, so these are not necessarily the "correct" answers; they are just possible answers.

1) What destination received most flights in June?

```
flights %>% filter(month==6) %>% count(dest)
     %>% arrange(n) %>% top_n(1)

# 1   ORD  1547
```

2) Which carrier had the greatest average distance per flight?

```
flights %>% group_by(carrier) %>%
     summarize(avg_dist=mean(distance)) %>%
     top_n(1, avg_dist) %>% inner_join(airlines)
# 1      HA     4983 Hawaiian Airlines Inc.
```

3) Which flight traveled the fastest?

```
 flights %>% mutate(speed=distance/(air_time/60)) %>%
     top_n(10, speed) %>%
     select(flight, dest, distance, speed)

#   flight  dest distance    speed
# 1   1499   ATL      762 703.3846
```

4) What day had the largest average arrival delay for all flights?

```
flights %>%
  filter(arr_delay > 0, !is.na(arr_delay)) %>%
  group_by(month, day) %>%
  summarise(avg_delay =  mean(arr_delay)) %>%
  ungroup() %>%
  top_n(1, avg_delay)
#   month   day avg_delay
#   <int> <int>     <dbl>
# 1     7    10      110.
```

5) What was the total distance for all flights in January? What was the average distance per flight?

```
flights %>% filter(month==1) %>%
      summarize(total=sum(distance), avg=mean(distance))
#       total      avg
# 1 27188805 1006.844
```

6) What day of the week saw the most flights?

```
flights %>% count(weekdays(time_hour)) %>% top_n(1, n)
# 1              Monday 50690
```

7) What was the average number of seats and engines on the plains that left from NYC on July 4?

```
flights %>% filter(month==7, day==4) %>%
      inner_join(planes, "tailnum") %>%
      select(seats, engines) %>%
      summarize_all(mean)
#       seats  engines
#       <dbl>    <dbl>
# 1 140.6581 1.991974
```

8) How many airlines do not have the word "air" somewhere in their name?

```
airlines %>% filter(!grepl("air", name, ignore.case=T)) %>%
      count()
airlines %>% filter(!str_detect(name, fixed("air",
ignore_case=T))) %>%
               count()
# just 1, Virgin America
```

9) What was the most common plane model to fly out of NYC in October?

```
flights %>% filter(month==10) %>%
      inner_join(planes, "tailnum") %>%
      count(model) %>%
      top_n(1,n)
# 1 A320-232  3717
```

10) How many planes (tailnum) only flew one route (flight) but flew that route more than 10 times?

```
flights %>%
  group_by(tailnum) %>%
  summarize(routes = n_distinct(flight), flights=n()) %>%
  filter(routes==1, flights>10) %>% nrow()
# [1] 12
```

11) Which scheduled departure hour (use the "hour" column) had the largest proportion of flights delayed (dep_delay) longer than 5 min?

```
flights %>% group_by(hour) %>%
  summarize(perc_delay=mean(dep_delay>5, na.rm=T)) %>%
  top_n(1, perc_delay)

#  hour perc_delay
#   <dbl>      <dbl>
# 1    21      0.463
```

12) Which flight(s) had the greatest scheduled length (time between scheduled departure and arrival)?

```
# helper function because times are weird
# 600 is 6:00 and 500 is 5:00 so
# timediff(600,500) should be 60 minutes, not 100
flights %>%
  mutate(
    arr_hour = sched_arr_time %/% 100,
    arr_min = sched_arr_time %% 100,
    arr_hms = arr_hour * 60 + arr_min,
    dep_hour = sched_dep_time %/% 100,
    dep_min = sched_dep_time %% 100,
    dep_hms = dep_hour * 60 + dep_min,
    sch_length=arr_hms-dep_hms + if_else(arr_hms < dep_hms,
24*60,0)
  ) %>%
  arrange(desc(sch_length)) %>%
  distinct(flight, sch_length) %>%
  top_n(1, sch_length)

# 1     51    400
```