Data Analysis Lab

Assignment Instructions

Complete all questions below. After completing the assignment, knit your document, and download both your .Rmd and knitted output. Upload your files for peer review.

For each response, include comments detailing your response and what each line does.

```
library(nycflights13)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Question 1.

Using the nycflights13 dataset, find all flights that departed in July, August, or September using the helper function between().

```
flights_jul_aug_sep <- flights %>%
  filter(between(month, 7, 9))

flights_jul_aug_sep
```

y… <int>	mo… <int>	da <int>	dep_time <int>	sched_dep_time <int>	dep_delay <dbl>	arr_time <int>	sched_arr_time <int>	arr_delay <dbl>
2013	7	1	1	2029	212	236	2359	157
2013	7	1	2	2359	3	344	344	0
2013	7	1	29	2245	104	151	1	110
2013	7	1	43	2130	193	322	14	188
2013	7	1	44	2150	174	300	100	120
2013	7	1	46	2051	235	304	2358	186
2013	7	1	48	2001	287	308	2305	243
2013	7	1	58	2155	183	335	43	172

y…	mo…	da	dep_time	sched_dep_time	dep_delay	arr_time	sched_arr_time	arr_delay
<int>	<int>	<int>	<int>	<int>	<dbl>	<int>	<int>	<dbl>
2013	7	1	100	2146	194	327	30	177
2013	7	1	100	2245	135	337	135	122

1-10 of 10,000 rows | 1-10 of 19 columns Previous **1** 2 3 4 5 6 … 1000 Next

Question 2.

Using the nycflights13 dataset sort flights to find the 10 flights that flew the furthest. Put them in order of fastest to slowest.

```
longest_flights <- flights %>%
  arrange(desc(distance)) %>%
  head(10) %>%
  mutate(speed = distance / (air_time / 60)) %>%  # calculate speed in miles per hour
  arrange(desc(speed))

longest_flights
```

y…	mo…	da	dep_time	sched_dep_time	dep_delay	arr_time	sched_arr_time	arr_delay
<int>	<int>	<int>	<int>	<int>	<dbl>	<int>	<int>	<dbl>
2013	1	6	1019	900	79	1558	1530	28
2013	1	7	1042	900	102	1620	1530	50
2013	1	3	914	900	14	1504	1530	-26
2013	1	10	859	900	-1	1449	1530	-41
2013	1	5	858	900	-2	1519	1530	-11
2013	1	2	909	900	9	1525	1530	-5
2013	1	4	900	900	0	1516	1530	-14
2013	1	9	641	900	1301	1242	1530	1272
2013	1	8	901	900	1	1504	1530	-26
2013	1	1	857	900	-3	1516	1530	-14

1-10 of 10 rows | 1-10 of 20 columns

Question 3.

Using the nycflights13 dataset, calculate a new variable called “hr_delay” and arrange the flights dataset in order of the arrival delays in hours (longest delays at the top). Put the new variable you created just before the departure time.Hint: use the experimental argument .before.

```
flights_with_hr_delay <- flights %>%
  mutate(hr_delay = arr_delay / 60) %>%
  arrange(desc(hr_delay)) %>%
  select(year:day, hr_delay, dep_time, everything())

flights_with_hr_delay
```

y… <int>	mo… <int>	da <int>	hr_delay <dbl>	dep_time <int>	sched_dep_time <int>	dep_delay <dbl>	arr_time <int>	sched_arr_time <int>
2013	1	9	21.200000	641	900	1301	1242	1530
2013	6	15	18.783333	1432	1935	1137	1607	2120
2013	1	10	18.483333	1121	1635	1126	1239	1810
2013	9	20	16.783333	1139	1845	1014	1457	2210
2013	7	22	16.483333	845	1600	1005	1044	1815
2013	4	10	15.516667	1100	1900	960	1342	2211
2013	3	17	15.250000	2321	810	911	135	1020
2013	7	22	14.916667	2257	759	898	121	1026
2013	12	5	14.633333	756	1700	896	1058	2020
2013	5	3	14.583333	1133	2055	878	1250	2215

1-10 of 10,000 rows | 1-10 of 20 columns Previous **1** 2 3 4 5 6 … 1000 Next

Question 4.

Using the nycflights13 dataset, find the most popular destinations (those with more than 2000 flights) and show the destination, the date info, the carrier. Then show just the number of flights for each popular destination.

```
popular_destinations <- flights %>%
  group_by(dest) %>%
  filter(n() > 2000) %>%
  select(dest, year, month, day, carrier)

number_of_flights <- popular_destinations %>%
  group_by(dest) %>%
  summarise(num_flights = n())

list(popular_destinations, number_of_flights)
```

```
## [[1]]
## # A tibble: 302,969 × 5
## # Groups:   dest [46]
##    dest   year month   day carrier
##    <chr> <int> <int> <int> <chr>
##  1 IAH    2013     1     1 UA
##  2 IAH    2013     1     1 UA
##  3 MIA    2013     1     1 AA
##  4 ATL    2013     1     1 DL
##  5 ORD    2013     1     1 UA
##  6 FLL    2013     1     1 B6
##  7 IAD    2013     1     1 EV
##  8 MCO    2013     1     1 B6
##  9 ORD    2013     1     1 AA
## 10 PBI    2013     1     1 B6
## # ℹ 302,959 more rows
##
## [[2]]
## # A tibble: 46 × 2
##    dest  num_flights
##    <chr>       <int>
##  1 ATL         17215
##  2 AUS          2439
##  3 BNA          6333
##  4 BOS         15508
##  5 BTV          2589
##  6 BUF          4681
##  7 CHS          2884
##  8 CLE          4573
##  9 CLT         14064
## 10 CMH          3524
## # ℹ 36 more rows
```

Question 5.

Using the nycflights13 dataset, find the flight information (flight number, origin, destination, carrier, number of flights in the year, and percent late) for the flight numbers with the highest percentage of arrival delays. Only include the flight numbers that have over 100 flights in the year.

```
flight_info <- flights %>%
  group_by(flight) %>%
  filter(n() > 100) %>%
  summarise(
    origin = first(origin),
    dest = first(dest),
    carrier = first(carrier),
    num_flights = n(),
    percent_late = mean(arr_delay > 0) * 100
  ) %>%
  arrange(desc(percent_late))

flight_info
```

flight <int>	origin <chr>	dest <chr>	carrier <chr>	num_flights <int>	percent_late <dbl>
43	JFK	MCO	B6	133	63.15789
803	JFK	SJU	B6	142	61.26761
1165	EWR	LAX	UA	147	61.22449
1127	EWR	SFO	UA	107	58.87850
195	EWR	MDW	WN	142	58.45070
705	JFK	SJU	B6	225	53.77778
1202	EWR	MIA	UA	106	53.77358
137	JFK	RSW	B6	153	53.59477
141	JFK	PBI	B6	377	52.78515
1130	LGA	IAH	UA	103	49.51456