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.

Question 1. Using the nycflights13 dataset, find all flights that departed in July, August, or September using the helper function between().

```
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
```

```
flights %>%
  filter(between(month, 7, 9)) %>%
  select(year, month, day, dep_time, origin, dest)
```

```
## # A tibble: 86,326 x 6
##     year month   day dep_time origin dest
##    <int> <int> <int>    <int> <chr>  <chr>
##  1  2013     7     1        1 JFK    SFO
##  2  2013     7     1        2 JFK    SJU
##  3  2013     7     1       29 JFK    BTV
##  4  2013     7     1       43 LGA    FLL
##  5  2013     7     1       44 JFK    LAX
##  6  2013     7     1       46 JFK    PDX
##  7  2013     7     1       48 JFK    LAX
##  8  2013     7     1       58 JFK    TPA
##  9  2013     7     1      100 JFK    MCO
## 10  2013     7     1      100 JFK    LAX
## # ... with 86,316 more rows
```

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.

```
library(nycflights13)

# Create a new column with the distance in miles
flights$distance_miles <- flights$distance * 0.621371

# Sort the flights by distance in descending order and select the top 10
top_10_flights <- head(flights[order(-flights$distance_miles),], 10)

# Sort the top 10 flights by air_time in ascending order
top_10_flights <- top_10_flights[order(top_10_flights$air_time),]

# Print the results
top_10_flights[, c("year", "month", "day", "carrier", "flight", "tailnum", "distance_miles", "air_time"]
```

```
## # A tibble: 10 x 8
##     year month   day carrier flight tailnum distance_miles air_time
##    <int> <int> <int> <chr>    <int> <chr>            <dbl>    <dbl>
##  1  2013     1     6 HA          51 N385HA           3096.      611
##  2  2013     1     7 HA          51 N385HA           3096.      612
##  3  2013     1     3 HA          51 N380HA           3096.      616
##  4  2013     1    10 HA          51 N388HA           3096.      633
##  5  2013     1     5 HA          51 N381HA           3096.      635
##  6  2013     1     2 HA          51 N380HA           3096.      638
##  7  2013     1     4 HA          51 N384HA           3096.      639
##  8  2013     1     9 HA          51 N384HA           3096.      640
##  9  2013     1     8 HA          51 N389HA           3096.      645
## 10  2013     1     1 HA          51 N380HA           3096.      659
```

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.

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

flights <- nycflights13::flights

# Calculate the arrival delay in hours
flights$hr_delay <- flights$arr_delay / 60

# Sort the dataset by hr_delay in descending order
flights <- flights %>% arrange(desc(hr_delay))

# Move the hr_delay column before the departure_time column
flights <- flights %>% select(hr_delay, everything())

#
```

```r
# Print the first 10 rows of the sorted and updated dataset
head(flights, 10)
```

```
## # A tibble: 10 x 20
##    hr_delay  year month   day dep_time sched_d~1 dep_d~2 arr_t~3 sched~4 arr_d~5
##       <dbl> <int> <int> <int>    <int>     <int>   <dbl>   <int>   <int>   <dbl>
##  1     21.2  2013     1     9      641       900    1301    1242    1530    1272
##  2     18.8  2013     6    15     1432      1935    1137    1607    2120    1127
##  3     18.5  2013     1    10     1121      1635    1126    1239    1810    1109
##  4     16.8  2013     9    20     1139      1845    1014    1457    2210    1007
##  5     16.5  2013     7    22      845      1600    1005    1044    1815     989
##  6     15.5  2013     4    10     1100      1900     960    1342    2211     931
##  7     15.2  2013     3    17     2321       810     911     135    1020     915
##  8     14.9  2013     7    22     2257       759     898     121    1026     895
##  9     14.6  2013    12     5      756      1700     896    1058    2020     878
## 10     14.6  2013     5     3     1133      2055     878    1250    2215     875
## # ... with 10 more variables: carrier <chr>, flight <int>, tailnum <chr>,
## #   origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>, hour <dbl>,
## #   minute <dbl>, time_hour <dttm>, and abbreviated variable names
## #   1: sched_dep_time, 2: dep_delay, 3: arr_time, 4: sched_arr_time,
## #   5: arr_delay
```

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.

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

flights <- nycflights13::flights

# Find the most popular destinations with more than 2000 flights
popular_destinations <- flights %>%
  group_by(dest) %>%
  filter(n() > 2000) %>%
  summarise(total_flights = n())

# Join the popular destinations with the original dataset to get the desired columns
popular_flights <- flights %>%
  inner_join(popular_destinations, by = "dest") %>%
  select(dest, year, month, day, carrier, total_flights)

# Show the number of flights for each popular destination
popular_flights %>% count(dest, sort = TRUE)
```

```
## # A tibble: 46 x 2
##    dest      n
##    <chr> <int>
##  1 ORD   17283
##  2 ATL   17215
##  3 LAX   16174
##  4 BOS   15508
```

```
##  5 MCO   14082
##  6 CLT   14064
##  7 SFO   13331
##  8 FLL   12055
##  9 MIA   11728
## 10 DCA    9705
## # ... with 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.

```
library(nycflights13)
library(dplyr)

flights <- nycflights13::flights

# Create a new variable that indicates whether the flight was delayed or not
flights <- flights %>%
  mutate(delayed = arr_delay > 0)

# Group flights by flight number, origin, destination, and carrier, and calculate the number of flights
flight_summary <- flights %>%
  group_by(flight, origin, dest, carrier) %>%
  filter(n() > 100) %>%
  summarise(num_flights = n(),
            percent_late = mean(delayed) * 100)
```

```
## `summarise()` has grouped output by 'flight', 'origin', 'dest'. You can
## override using the `.groups` argument.
```

```
# Arrange the flight summary by the percentage of late flights in descending order
flight_summary <- flight_summary %>%
  arrange(desc(percent_late))

# Join the flight summary with the original dataset to get the desired columns
flights_info <- flights %>%
  inner_join(flight_summary, by = c("flight", "origin", "dest", "carrier")) %>%
  select(flight, origin, dest, carrier, num_flights, percent_late)

# Show the flights with the highest percentage of late flights
head(flights_info, 10)
```

```
## # A tibble: 10 x 6
##    flight origin dest  carrier num_flights percent_late
##     <int> <chr>  <chr> <chr>         <int>        <dbl>
##  1   1714 LGA    IAH   UA              140         34.3
##  2    725 JFK    BQN   B6              119         NA
##  3    461 LGA    ATL   DL              268         NA
##  4    507 EWR    FLL   B6              168         NA
##  5     79 JFK    MCO   B6              162         NA
##  6    301 LGA    ORD   AA              285         NA
```

```
##  7    707 LGA    DFW   AA             242         NA
##  8    371 LGA    FLL   B6             365         NA
##  9   4650 LGA    ATL   MQ             116         NA
## 10   1919 LGA    MSP   DL             326         NA
```