Stat 261, Lab 7, Solutions

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

In this lab we will work with the `nycflights13` data.

1. Add the latitude and longitude of each airport destination to the `flights` table using a `join` function. You will find the data on latitude and longitude in the `airports` table.

```r
flights %>%
  left_join(select(airports,faa,lon,lat),by=c("dest"="faa"))
```

```
## # A tibble: 336,776 x 21
##     year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##    <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
##  1  2013     1     1      517            515         2      830            819
##  2  2013     1     1      533            529         4      850            830
##  3  2013     1     1      542            540         2      923            850
##  4  2013     1     1      544            545        -1     1004           1022
##  5  2013     1     1      554            600        -6      812            837
##  6  2013     1     1      554            558        -4      740            728
##  7  2013     1     1      555            600        -5      913            854
##  8  2013     1     1      557            600        -3      709            723
##  9  2013     1     1      557            600        -3      838            846
## 10  2013     1     1      558            600        -2      753            745
## # ... with 336,766 more rows, and 13 more variables: arr_delay <dbl>,
## #   carrier <chr>, flight <int>, tailnum <chr>, origin <chr>, dest <chr>,
## #   air_time <dbl>, distance <dbl>, hour <dbl>, minute <dbl>, time_hour <dttm>,
## #   lon <dbl>, lat <dbl>
```

2. Create a table with the year-month-day-flight-tailnum combinations that have more than 1 flight (careful about missing tailnum). Use this table to filter the flights table and then select carrire, flight, origin and dest. Which airline used the same flight number for a plane that made a trip from La Guardia to St. Louis in the morning and from Newark to Denver in the afternoon?

```r
tem <- flights %>%
  count(year,month,day,flight,tailnum) %>%
  filter(n>1,!is.na(tailnum))
flights %>% semi_join(tem) %>%
  select(year:day,carrier,flight,origin,dest)
```

```
## # A tibble: 14 x 7
##     year month   day carrier flight origin dest
##    <int> <int> <int> <chr>    <int> <chr>  <chr>
##  1  2013     6     8 WN        2269 LGA    STL
##  2  2013     6     8 WN        2269 EWR    DEN
##  3  2013     6    15 WN        2269 LGA    STL
##  4  2013     6    15 WN        2269 EWR    DEN
##  5  2013     6    22 WN        2269 LGA    STL
##  6  2013     6    22 WN        2269 EWR    DEN
##  7  2013     6    29 WN        2269 LGA    STL
##  8  2013     6    29 WN        2269 EWR    DEN
##  9  2013     7     6 WN        2269 LGA    STL
## 10  2013     7     6 WN        2269 EWR    DEN
## 11  2013     8     3 WN        2269 LGA    STL
## 12  2013     8     3 WN        2269 EWR    DEN
## 13  2013     8    10 WN        2269 LGA    STL
## 14  2013     8    10 WN        2269 EWR    DEN
```

```
# WN = Southwest used the same flight number for a plane that made a trip from La Guardia
# to St. Louis in the morning and from Newark to Denver in the afternoon.
```

3. One of the exercises in the lecture 7 notes asked you to create a table called `top_dep_delay` from the flights table. `top_dep_delay` was comprised of the year-month-days with the 3 largest total delays, where total delay is defined as the sum of the `dep_delay` variable for each year-month-day. Recreate `top_dep_delay` for this lab exercise. For each of the three top-delay days, report the median, third quartile and maximum of the dep_delay variable in `flights`.

```
top_dep_delay <- flights %>%
  group_by(year,month,day) %>%
  summarize(tot_delay = sum(dep_delay,na.rm=TRUE)) %>%
  arrange(desc(tot_delay)) %>%
  head(3)
flights %>% semi_join(top_dep_delay) %>%
  group_by(year,month,day) %>%
  summarize(median=median(dep_delay,na.rm=TRUE),
            Q3=quantile(dep_delay,probs=.75,na.rm=TRUE),
            max=max(dep_delay,na.rm=TRUE))
```

```
## # A tibble: 3 x 6
## # Groups:   year, month [2]
##    year month   day median    Q3   max
##   <int> <int> <int>  <dbl> <dbl> <dbl>
## 1  2013     3     8     58  134.   470
## 2  2013     7     1     30    93   363
## 3  2013     7    10      7    69   634
```