HW4

Ayush

10/7/2020

Problem 1: This problem will involve the nycflights13 dataset (including tables airlines, airports, planes and weather), which we saw in class. Start by installing and importing the dataset to your chosen platform.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
library(nycflights13)
show(airports)
```

```
## # A tibble: 1,458 x 8
##    faa   name                      lat    lon   alt    tz dst   tzone
##    <chr> <chr>                   <dbl>  <dbl> <dbl> <dbl> <chr> <chr>
##  1 04G   Lansdowne Airport        41.1  -80.6  1044    -5 A     America/New_Yo~
##  2 06A   Moton Field Municipal A~ 32.5  -85.7   264    -6 A     America/Chicago
##  3 06C   Schaumburg Regional      42.0  -88.1   801    -6 A     America/Chicago
##  4 06N   Randall Airport          41.4  -74.4   523    -5 A     America/New_Yo~
##  5 09J   Jekyll Island Airport    31.1  -81.4    11    -5 A     America/New_Yo~
##  6 0A9   Elizabethton Municipal ~ 36.4  -82.2  1593    -5 A     America/New_Yo~
##  7 0G6   Williams County Airport  41.5  -84.5   730    -5 A     America/New_Yo~
##  8 0G7   Finger Lakes Regional A~ 42.9  -76.8   492    -5 A     America/New_Yo~
##  9 0P2   Shoestring Aviation Air~ 39.8  -76.6  1000    -5 U     America/New_Yo~
## 10 0S9   Jefferson County Intl    48.1 -123.    108    -8 A     America/Los_An~
## # ... with 1,448 more rows
```

```
flights
```

```
## # A tibble: 336,776 x 19
##     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 11 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>
```

a. (10 pts) Filter the dataset (using a left join) to display the tail number, year, month, day, hour, origin, and humidity for all flights heading to Tampa International Airport (TPA) on the afternoon of November 1, 2013.

```
data1 <- flights %>%
  select(year,month,day, hour,origin,tailnum,dest) %>%
  left_join(
    airports, by = c("dest"= "faa")) %>%
  left_join(weather) %>%
  filter(dest == "TPA", year == 2013, month == 11, day == 1, hour > 12) %>%
  select(tailnum, year, month, day, hour, origin, humid)
```

```
## Joining, by = c("year", "month", "day", "hour", "origin")
```

```
show(data1)
```

```
## # A tibble: 10 x 7
##    tailnum  year month   day  hour origin humid
##    <chr>   <int> <int> <int> <dbl> <chr>  <dbl>
##  1 N580JB   2013    11     1    14 JFK     63.1
##  2 N337NB   2013    11     1    14 LGA     56.5
##  3 N567UA   2013    11     1    15 EWR     52.8
##  4 N515MQ   2013    11     1    14 JFK     63.1
##  5 N779JB   2013    11     1    15 EWR     52.8
##  6 N561JB   2013    11     1    16 LGA     50.6
##  7 N974DL   2013    11     1    18 JFK     74.8
##  8 N319NB   2013    11     1    19 LGA     60.5
##  9 N76265   2013    11     1    19 EWR     72.5
## 10 N768JB   2013    11     1    19 JFK     83.5
```

b. (10 pts) What is the difference between the following two joins? anti_join(flights, airports, by = c("dest" = "faa")) anti_join(airports, flights, by = c("faa" = "dest"))

In the first anti_join , it will be returning only the flights that have gone to the airport that are NOT in the FAA list of destinations. Whereas the latter expression, will return the US airports that are just not the destination of any flight at all.

c. (10 pts) Select the origin and destination airports and their latitude and longitude for all fights in the dataset (using one or more inner joins). Hint: There should be 329,174 flights if you've done this correctly.

```
data2 <- flights %>%
            select(origin, dest) %>%
            left_join(airports, by = c("origin" = "faa")) %>%
            left_join(airports,by = c("dest" = "faa"),suffix = c("origin", "dest"))
data2
```

```
## # A tibble: 336,776 x 16
##    origin dest  nameorigin latorigin lonorigin altorigin tzorigin dstorigin
##    <chr>  <chr> <chr>          <dbl>     <dbl>     <dbl>    <dbl> <chr>
##  1 EWR    IAH   Newark Li~      40.7     -74.2        18       -5 A
##  2 LGA    IAH   La Guardia      40.8     -73.9        22       -5 A
##  3 JFK    MIA   John F Ke~      40.6     -73.8        13       -5 A
##  4 JFK    BQN   John F Ke~      40.6     -73.8        13       -5 A
##  5 LGA    ATL   La Guardia      40.8     -73.9        22       -5 A
##  6 EWR    ORD   Newark Li~      40.7     -74.2        18       -5 A
##  7 EWR    FLL   Newark Li~      40.7     -74.2        18       -5 A
##  8 LGA    IAD   La Guardia      40.8     -73.9        22       -5 A
##  9 JFK    MCO   John F Ke~      40.6     -73.8        13       -5 A
## 10 LGA    ORD   La Guardia      40.8     -73.9        22       -5 A
## # ... with 336,766 more rows, and 8 more variables: tzoneorigin <chr>,
## #   namedest <chr>, latdest <dbl>, londest <dbl>, altdest <dbl>, tzdest <dbl>,
## #   dstdest <chr>, tzonedest <chr>
```

d. (10 pts) Use group_by and count to get the number of flights to each unique origin/destination combination. Hint: There should be 217 of these total.

```
data3 <- flights %>%
  group_by(origin,dest) %>%
  count(sort = FALSE)
data3
```

```
## # A tibble: 224 x 3
## # Groups:   origin, dest [224]
##    origin dest      n
##    <chr>  <chr> <int>
##  1 EWR    ALB     439
##  2 EWR    ANC       8
##  3 EWR    ATL    5022
##  4 EWR    AUS     968
##  5 EWR    AVL     265
##  6 EWR    BDL     443
##  7 EWR    BNA    2336
##  8 EWR    BOS    5327
##  9 EWR    BQN     297
## 10 EWR    BTV     931
## # ... with 214 more rows
```

e. (10 pts) Produce a map that colors each destination airport by the average air time of its incoming flights. Here is a code snippet to draw a map of all flight destinations, which you can use as a starting

point. You may need to install the maps packages if you have not already. Adjust the title, axis labels and aesthetics to make this visualization as clear as possible. Hint: You may find it useful to use a different type of join in your solution than the one in the snippet.

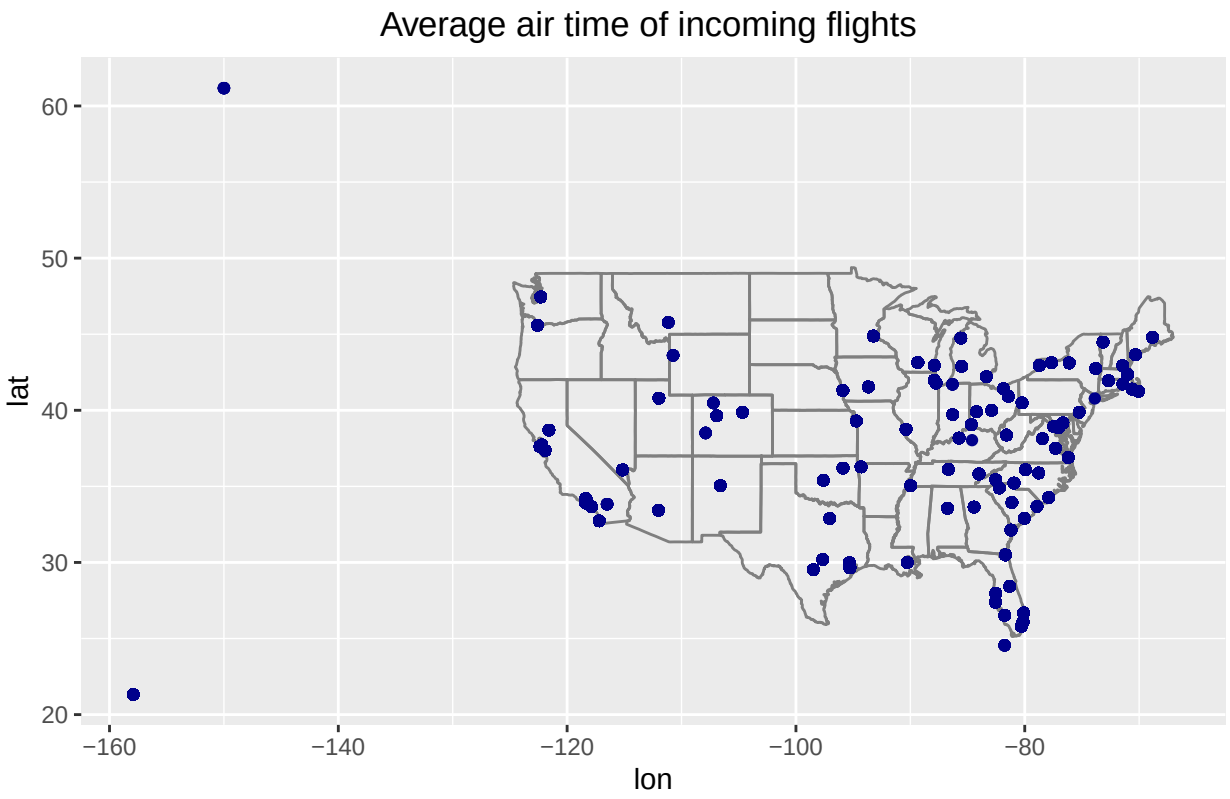
```
avg_air_time <- flights %>%
  group_by(dest) %>%
  summarise(avg_airtime = mean(arr_time, na.rm = TRUE)) %>%
  inner_join(airports, by = c(dest = "faa"))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
avg_air_time
```

```
## # A tibble: 101 x 9
##    dest  avg_airtime name                lat    lon   alt    tz dst   tzone
##    <chr>       <dbl> <chr>             <dbl>  <dbl> <dbl> <dbl> <chr> <chr>
##  1 ABQ         2049. Albuquerque Inte~  35.0 -107.   5355    -7 A     America/D~
##  2 ACK         1145. Nantucket Mem      41.3  -70.1    48    -5 A     America/N~
##  3 ALB         1702. Albany Intl        42.7  -73.8   285    -5 A     America/N~
##  4 ANC         1968  Ted Stevens Anch~  61.2 -150.    152    -9 A     America/A~
##  5 ATL         1513. Hartsfield Jacks~  33.6  -84.4  1026    -5 A     America/N~
##  6 AUS         1614. Austin Bergstrom~  30.2  -97.7   542    -6 A     America/C~
##  7 AVL         1373. Asheville Region~  35.4  -82.5  2165    -5 A     America/N~
##  8 BDL         1549. Bradley Intl       41.9  -72.7   173    -5 A     America/N~
##  9 BGR         1715. Bangor Intl        44.8  -68.8   192    -5 A     America/N~
## 10 BHM         2028. Birmingham Intl    33.6  -86.8   644    -6 A     America/C~
## # ... with 91 more rows
```

```
avg_air_time <- airports %>%
                inner_join(flights, c("faa" = "dest")) %>%
                ggplot(aes(lon, lat)) +
                borders("state") +
                geom_point(color = 'darkblue') +
                coord_quickmap()
avg_air_time <- avg_air_time + labs(title = "
                            Average air time of incoming flights")
avg_air_time
```


Average air time of incoming flights
lat
lon
60
50
40
30
20
−160
−140
−120
−100
−80