Stat. 450 Section 1 or 2: Homework 3

**Prof. Eric A. Suess**

So how should you complete your homework for this class?

* First thing to do is type all of your information about the problems you do in the text part of your R Notebook.
* Second thing to do is type all of your R code into R chunks that can be run.
* If you load the tidyverse in an R Notebook chunk, be sure to include the “message = FALSE” in the {r}, so {r message = FALSE}.
* Last thing is to spell check your R Notebook. Edit > Check Spelling… or hit the F7 key.

Homework 3:

Read: Chapter 5  
 Do 5.2.4 Exercises 1, 2,  
 Do 5.3.1 Exericise 2

library(tidyverse)

# 5.2.4

## 1.

This problem looks at the nycflights13 data set.

library(nycflights13)  
  
flights

## # A tibble: 336,776 x 19  
## year month day dep\_time sched\_dep\_time dep\_delay arr\_time  
## <int> <int> <int> <int> <int> <dbl> <int>  
## 1 2013 1 1 517 515 2 830  
## 2 2013 1 1 533 529 4 850  
## 3 2013 1 1 542 540 2 923  
## 4 2013 1 1 544 545 -1 1004  
## 5 2013 1 1 554 600 -6 812  
## 6 2013 1 1 554 558 -4 740  
## 7 2013 1 1 555 600 -5 913  
## 8 2013 1 1 557 600 -3 709  
## 9 2013 1 1 557 600 -3 838  
## 10 2013 1 1 558 600 -2 753  
## # ... with 336,766 more rows, and 12 more variables: sched\_arr\_time <int>,  
## # 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>

These questions relate to finding the flights that meet the conditions specified.

help(flights)

Note that dep\_delay is in minutes.

1. Had an arrival delay of two or more hours. So more than 2\*60 minutes.

flights %>% filter(dep\_delay >= 120)

## # A tibble: 9,888 x 19  
## year month day dep\_time sched\_dep\_time dep\_delay arr\_time  
## <int> <int> <int> <int> <int> <dbl> <int>  
## 1 2013 1 1 848 1835 853 1001  
## 2 2013 1 1 957 733 144 1056  
## 3 2013 1 1 1114 900 134 1447  
## 4 2013 1 1 1540 1338 122 2020  
## 5 2013 1 1 1815 1325 290 2120  
## 6 2013 1 1 1842 1422 260 1958  
## 7 2013 1 1 1856 1645 131 2212  
## 8 2013 1 1 1934 1725 129 2126  
## 9 2013 1 1 1938 1703 155 2109  
## 10 2013 1 1 1942 1705 157 2124  
## # ... with 9,878 more rows, and 12 more variables: sched\_arr\_time <int>,  
## # 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>

1. Flew to Houston (IAH or HOU)

flights %>% filter(dest == "IAH" | dest == "HOU")

## # A tibble: 9,313 x 19  
## year month day dep\_time sched\_dep\_time dep\_delay arr\_time  
## <int> <int> <int> <int> <int> <dbl> <int>  
## 1 2013 1 1 517 515 2 830  
## 2 2013 1 1 533 529 4 850  
## 3 2013 1 1 623 627 -4 933  
## 4 2013 1 1 728 732 -4 1041  
## 5 2013 1 1 739 739 0 1104  
## 6 2013 1 1 908 908 0 1228  
## 7 2013 1 1 1028 1026 2 1350  
## 8 2013 1 1 1044 1045 -1 1352  
## 9 2013 1 1 1114 900 134 1447  
## 10 2013 1 1 1205 1200 5 1503  
## # ... with 9,303 more rows, and 12 more variables: sched\_arr\_time <int>,  
## # 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>

or

flights %>% filter(dest %in% c("IAH", "HOU"))

## # A tibble: 9,313 x 19  
## year month day dep\_time sched\_dep\_time dep\_delay arr\_time  
## <int> <int> <int> <int> <int> <dbl> <int>  
## 1 2013 1 1 517 515 2 830  
## 2 2013 1 1 533 529 4 850  
## 3 2013 1 1 623 627 -4 933  
## 4 2013 1 1 728 732 -4 1041  
## 5 2013 1 1 739 739 0 1104  
## 6 2013 1 1 908 908 0 1228  
## 7 2013 1 1 1028 1026 2 1350  
## 8 2013 1 1 1044 1045 -1 1352  
## 9 2013 1 1 1114 900 134 1447  
## 10 2013 1 1 1205 1200 5 1503  
## # ... with 9,303 more rows, and 12 more variables: sched\_arr\_time <int>,  
## # 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>

1. Were operated by United, American, or Delta

Lets check the airlines dataframe for the codes for these airlines.

airlines

## # A tibble: 16 x 2  
## carrier name   
## <chr> <chr>   
## 1 9E Endeavor Air Inc.   
## 2 AA American Airlines Inc.   
## 3 AS Alaska Airlines Inc.   
## 4 B6 JetBlue Airways   
## 5 DL Delta Air Lines Inc.   
## 6 EV ExpressJet Airlines Inc.   
## 7 F9 Frontier Airlines Inc.   
## 8 FL AirTran Airways Corporation  
## 9 HA Hawaiian Airlines Inc.   
## 10 MQ Envoy Air   
## 11 OO SkyWest Airlines Inc.   
## 12 UA United Air Lines Inc.   
## 13 US US Airways Inc.   
## 14 VX Virgin America   
## 15 WN Southwest Airlines Co.   
## 16 YV Mesa Airlines Inc.

flights %>% filter(carrier %in% c("DL", "AA", "UA"))

## # A tibble: 139,504 x 19  
## year month day dep\_time sched\_dep\_time dep\_delay arr\_time  
## <int> <int> <int> <int> <int> <dbl> <int>  
## 1 2013 1 1 517 515 2 830  
## 2 2013 1 1 533 529 4 850  
## 3 2013 1 1 542 540 2 923  
## 4 2013 1 1 554 600 -6 812  
## 5 2013 1 1 554 558 -4 740  
## 6 2013 1 1 558 600 -2 753  
## 7 2013 1 1 558 600 -2 924  
## 8 2013 1 1 558 600 -2 923  
## 9 2013 1 1 559 600 -1 941  
## 10 2013 1 1 559 600 -1 854  
## # ... with 139,494 more rows, and 12 more variables: sched\_arr\_time <int>,  
## # 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>

1. Departed in summer (July, August, and September)

flights %>% filter(month %in% c(7,8,9))

## # A tibble: 86,326 x 19  
## year month day dep\_time sched\_dep\_time dep\_delay arr\_time  
## <int> <int> <int> <int> <int> <dbl> <int>  
## 1 2013 7 1 1 2029 212 236  
## 2 2013 7 1 2 2359 3 344  
## 3 2013 7 1 29 2245 104 151  
## 4 2013 7 1 43 2130 193 322  
## 5 2013 7 1 44 2150 174 300  
## 6 2013 7 1 46 2051 235 304  
## 7 2013 7 1 48 2001 287 308  
## 8 2013 7 1 58 2155 183 335  
## 9 2013 7 1 100 2146 194 327  
## 10 2013 7 1 100 2245 135 337  
## # ... with 86,316 more rows, and 12 more variables: sched\_arr\_time <int>,  
## # 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>

1. Arrived more than two hours late, but didn’t leave late

flights %>% filter(dep\_delay <= 0 & arr\_delay > 2\*60)

## # A tibble: 29 x 19  
## year month day dep\_time sched\_dep\_time dep\_delay arr\_time  
## <int> <int> <int> <int> <int> <dbl> <int>  
## 1 2013 1 27 1419 1420 -1 1754  
## 2 2013 10 7 1350 1350 0 1736  
## 3 2013 10 7 1357 1359 -2 1858  
## 4 2013 10 16 657 700 -3 1258  
## 5 2013 11 1 658 700 -2 1329  
## 6 2013 3 18 1844 1847 -3 39  
## 7 2013 4 17 1635 1640 -5 2049  
## 8 2013 4 18 558 600 -2 1149  
## 9 2013 4 18 655 700 -5 1213  
## 10 2013 5 22 1827 1830 -3 2217  
## # ... with 19 more rows, and 12 more variables: sched\_arr\_time <int>,  
## # 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>

1. Were delayed by at least an hour, but made up over 30 minutes in flight

flights %>% filter(dep\_delay >= 60 & dep\_delay - arr\_delay > 30 )

## # A tibble: 1,844 x 19  
## year month day dep\_time sched\_dep\_time dep\_delay arr\_time  
## <int> <int> <int> <int> <int> <dbl> <int>  
## 1 2013 1 1 2205 1720 285 46  
## 2 2013 1 1 2326 2130 116 131  
## 3 2013 1 3 1503 1221 162 1803  
## 4 2013 1 3 1839 1700 99 2056  
## 5 2013 1 3 1850 1745 65 2148  
## 6 2013 1 3 1941 1759 102 2246  
## 7 2013 1 3 1950 1845 65 2228  
## 8 2013 1 3 2015 1915 60 2135  
## 9 2013 1 3 2257 2000 177 45  
## 10 2013 1 4 1917 1700 137 2135  
## # ... with 1,834 more rows, and 12 more variables: sched\_arr\_time <int>,  
## # 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>

1. Departed between midnight and 6am (inclusive)

Note that midnight is 2400, not 0.

flights %>% filter(dep\_time <= 600 | dep\_time == 2400 )

## # A tibble: 9,373 x 19  
## year month day dep\_time sched\_dep\_time dep\_delay arr\_time  
## <int> <int> <int> <int> <int> <dbl> <int>  
## 1 2013 1 1 517 515 2 830  
## 2 2013 1 1 533 529 4 850  
## 3 2013 1 1 542 540 2 923  
## 4 2013 1 1 544 545 -1 1004  
## 5 2013 1 1 554 600 -6 812  
## 6 2013 1 1 554 558 -4 740  
## 7 2013 1 1 555 600 -5 913  
## 8 2013 1 1 557 600 -3 709  
## 9 2013 1 1 557 600 -3 838  
## 10 2013 1 1 558 600 -2 753  
## # ... with 9,363 more rows, and 12 more variables: sched\_arr\_time <int>,  
## # 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>

## 2.

The between() function can be used like %in%.

The between() can be used with the months to filter the rows from July, August, September.

flights %>% filter(between(month, 7, 9))

## # A tibble: 86,326 x 19  
## year month day dep\_time sched\_dep\_time dep\_delay arr\_time  
## <int> <int> <int> <int> <int> <dbl> <int>  
## 1 2013 7 1 1 2029 212 236  
## 2 2013 7 1 2 2359 3 344  
## 3 2013 7 1 29 2245 104 151  
## 4 2013 7 1 43 2130 193 322  
## 5 2013 7 1 44 2150 174 300  
## 6 2013 7 1 46 2051 235 304  
## 7 2013 7 1 48 2001 287 308  
## 8 2013 7 1 58 2155 183 335  
## 9 2013 7 1 100 2146 194 327  
## 10 2013 7 1 100 2245 135 337  
## # ... with 86,316 more rows, and 12 more variables: sched\_arr\_time <int>,  
## # 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>

# 5.3.1

# 2.

Sort flights to find the most delayed flights. Find the flights that left earliest.

The five most delayed flights.

flights %>% arrange(desc(dep\_delay)) %>%  
 head(5)

## # A tibble: 5 x 19  
## year month day dep\_time sched\_dep\_time dep\_delay arr\_time  
## <int> <int> <int> <int> <int> <dbl> <int>  
## 1 2013 1 9 641 900 1301 1242  
## 2 2013 6 15 1432 1935 1137 1607  
## 3 2013 1 10 1121 1635 1126 1239  
## 4 2013 9 20 1139 1845 1014 1457  
## 5 2013 7 22 845 1600 1005 1044  
## # ... with 12 more variables: sched\_arr\_time <int>, 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>

The five flights that left the earliest.

flights %>% arrange(dep\_delay) %>%  
 head(5)

## # A tibble: 5 x 19  
## year month day dep\_time sched\_dep\_time dep\_delay arr\_time  
## <int> <int> <int> <int> <int> <dbl> <int>  
## 1 2013 12 7 2040 2123 -43 40  
## 2 2013 2 3 2022 2055 -33 2240  
## 3 2013 11 10 1408 1440 -32 1549  
## 4 2013 1 11 1900 1930 -30 2233  
## 5 2013 1 29 1703 1730 -27 1947  
## # ... with 12 more variables: sched\_arr\_time <int>, 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>