Stat. 450 Section 1 or 2: Homework 9

**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.

Upload one file to Blackboard.

Homework 9:

 Read: Chapter 13

 Exercises:

 Do 13.2.1 Exercises 1, 3
 Do 13.3.1 Exercise 1
 Do 13.4.6 Exercises 1, 2, 3

# 13.2.1

## 1. Imagine you wanted to draw (approximately) the route each plane flies from its origin to its destination. What variables would you need? What tables would you need to combine?

**Answer:** Need flights and airports. From flights get origin and dest. From airports get lat and long.

library(tidyverse)

## ── Attaching packages ───────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────── tidyverse 1.2.1 ──

## ✔ ggplot2 3.1.0 ✔ purrr 0.2.5
## ✔ tibble 1.4.2 ✔ dplyr 0.7.7
## ✔ tidyr 0.8.2 ✔ stringr 1.3.1
## ✔ readr 1.1.1 ✔ forcats 0.3.0

## ── Conflicts ──────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────── tidyverse\_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag() masks stats::lag()

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>

airports

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

flights <- flights %>% left\_join(airports, c("origin" = "faa"))

flights <- flights %>% left\_join(airports, c("dest" = "faa"))

flights

## # A tibble: 336,776 x 33
## 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 26 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>, name.x <chr>, lat.x <dbl>,
## # lon.x <dbl>, alt.x <int>, tz.x <dbl>, dst.x <chr>, tzone.x <chr>,
## # name.y <chr>, lat.y <dbl>, lon.y <dbl>, alt.y <int>, tz.y <dbl>,
## # dst.y <chr>, tzone.y <chr>

## 3. weather only contains information for the origin (NYC) airports. If it contained weather records for all airports in the USA, what additional relation would it define with flights?

**Answer:** If all airports were included then the weather at the destination would be available also. Note that the year, month, day, hour would be used for the destination location’s weather.

weather

## # A tibble: 26,115 x 15
## origin year month day hour temp dewp humid wind\_dir wind\_speed
## <chr> <dbl> <dbl> <int> <int> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 EWR 2013 1 1 1 39.0 26.1 59.4 270 10.4
## 2 EWR 2013 1 1 2 39.0 27.0 61.6 250 8.06
## 3 EWR 2013 1 1 3 39.0 28.0 64.4 240 11.5
## 4 EWR 2013 1 1 4 39.9 28.0 62.2 250 12.7
## 5 EWR 2013 1 1 5 39.0 28.0 64.4 260 12.7
## 6 EWR 2013 1 1 6 37.9 28.0 67.2 240 11.5
## 7 EWR 2013 1 1 7 39.0 28.0 64.4 240 15.0
## 8 EWR 2013 1 1 8 39.9 28.0 62.2 250 10.4
## 9 EWR 2013 1 1 9 39.9 28.0 62.2 260 15.0
## 10 EWR 2013 1 1 10 41 28.0 59.6 260 13.8
## # ... with 26,105 more rows, and 5 more variables: wind\_gust <dbl>,
## # precip <dbl>, pressure <dbl>, visib <dbl>, time\_hour <dttm>

flights

## # A tibble: 336,776 x 33
## 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 26 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>, name.x <chr>, lat.x <dbl>,
## # lon.x <dbl>, alt.x <int>, tz.x <dbl>, dst.x <chr>, tzone.x <chr>,
## # name.y <chr>, lat.y <dbl>, lon.y <dbl>, alt.y <int>, tz.y <dbl>,
## # dst.y <chr>, tzone.y <chr>

# 13.3.1.

## 1. Add a surrogate key to flights.

flights

## # A tibble: 336,776 x 33
## 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 26 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>, name.x <chr>, lat.x <dbl>,
## # lon.x <dbl>, alt.x <int>, tz.x <dbl>, dst.x <chr>, tzone.x <chr>,
## # name.y <chr>, lat.y <dbl>, lon.y <dbl>, alt.y <int>, tz.y <dbl>,
## # dst.y <chr>, tzone.y <chr>

flights %>% mutate(index = row\_number())

## # A tibble: 336,776 x 34
## 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 27 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>, name.x <chr>, lat.x <dbl>,
## # lon.x <dbl>, alt.x <int>, tz.x <dbl>, dst.x <chr>, tzone.x <chr>,
## # name.y <chr>, lat.y <dbl>, lon.y <dbl>, alt.y <int>, tz.y <dbl>,
## # dst.y <chr>, tzone.y <chr>, index <int>

# 13.4.6

## 1. Compute the average delay by destination, then join on the airports data frame so you can show the spatial distribution of delays. Here’s an easy way to draw a map of the United States:

flights

## # A tibble: 336,776 x 33
## 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 26 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>, name.x <chr>, lat.x <dbl>,
## # lon.x <dbl>, alt.x <int>, tz.x <dbl>, dst.x <chr>, tzone.x <chr>,
## # name.y <chr>, lat.y <dbl>, lon.y <dbl>, alt.y <int>, tz.y <dbl>,
## # dst.y <chr>, tzone.y <chr>

delays <- flights %>% group\_by(dest) %>%
 summarise(delay\_ave = mean(arr\_delay, na.rm = TRUE))
delays

## # A tibble: 105 x 2
## dest delay\_ave
## <chr> <dbl>
## 1 ABQ 4.38
## 2 ACK 4.85
## 3 ALB 14.4
## 4 ANC -2.5
## 5 ATL 11.3
## 6 AUS 6.02
## 7 AVL 8.00
## 8 BDL 7.05
## 9 BGR 8.03
## 10 BHM 16.9
## # ... with 95 more rows

airports

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

delays <- delays %>% inner\_join(airports, by = c("dest" = "faa"))
delays

## # A tibble: 101 x 9
## dest delay\_ave name lat lon alt tz dst tzone
## <chr> <dbl> <chr> <dbl> <dbl> <int> <dbl> <chr> <chr>
## 1 ABQ 4.38 Albuquerque In… 35.0 -107. 5355 -7 A America…
## 2 ACK 4.85 Nantucket Mem 41.3 -70.1 48 -5 A America…
## 3 ALB 14.4 Albany Intl 42.7 -73.8 285 -5 A America…
## 4 ANC -2.5 Ted Stevens An… 61.2 -150. 152 -9 A America…
## 5 ATL 11.3 Hartsfield Jac… 33.6 -84.4 1026 -5 A America…
## 6 AUS 6.02 Austin Bergstr… 30.2 -97.7 542 -6 A America…
## 7 AVL 8.00 Asheville Regi… 35.4 -82.5 2165 -5 A America…
## 8 BDL 7.05 Bradley Intl 41.9 -72.7 173 -5 A America…
## 9 BGR 8.03 Bangor Intl 44.8 -68.8 192 -5 A America…
## 10 BHM 16.9 Birmingham Intl 33.6 -86.8 644 -6 A America…
## # ... with 91 more rows

delays %>%
 ggplot(aes(lon, lat, color = delay\_ave)) +
 borders("state") +
 geom\_point() +
 coord\_quickmap()

##
## Attaching package: 'maps'

## The following object is masked from 'package:purrr':
##
## map



## 2. Add the location of the origin and destination (i.e. the lat and lon) to flights.

airports

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

airports\_loc <- airports %>%
 select(faa, lat, lon)

flights %>%
 select(year:day, hour, origin, dest) %>%
 left\_join(
 airports\_loc,
 by = c("origin" = "faa")
 ) %>%
 left\_join(
 airports\_loc,
 by = c("dest" = "faa")
 )

## # A tibble: 336,776 x 10
## year month day hour origin dest lat.x lon.x lat.y lon.y
## <int> <int> <int> <dbl> <chr> <chr> <dbl> <dbl> <dbl> <dbl>
## 1 2013 1 1 5 EWR IAH 40.7 -74.2 30.0 -95.3
## 2 2013 1 1 5 LGA IAH 40.8 -73.9 30.0 -95.3
## 3 2013 1 1 5 JFK MIA 40.6 -73.8 25.8 -80.3
## 4 2013 1 1 5 JFK BQN 40.6 -73.8 NA NA
## 5 2013 1 1 6 LGA ATL 40.8 -73.9 33.6 -84.4
## 6 2013 1 1 5 EWR ORD 40.7 -74.2 42.0 -87.9
## 7 2013 1 1 6 EWR FLL 40.7 -74.2 26.1 -80.2
## 8 2013 1 1 6 LGA IAD 40.8 -73.9 38.9 -77.5
## 9 2013 1 1 6 JFK MCO 40.6 -73.8 28.4 -81.3
## 10 2013 1 1 6 LGA ORD 40.8 -73.9 42.0 -87.9
## # ... with 336,766 more rows

## 3. Is there a relationship between the age of a plane and its delays?

plane\_ages <-
 planes %>%
 mutate(age = 2013 - year) %>%
 select(tailnum, age)

flights %>%
 inner\_join(plane\_ages, by = "tailnum") %>%
 group\_by(age) %>%
 filter(!is.na(dep\_delay)) %>%
 summarise(delay = mean(dep\_delay)) %>%
 ggplot(aes(x = age, y = delay)) +
 geom\_point() +
 geom\_line()

## Warning: Removed 1 rows containing missing values (geom\_point).

## Warning: Removed 1 rows containing missing values (geom\_path).

