Assingment NYC Flight Data

Thuyen Nguyen

April 10, 2017

NYC Flight Data

Your job is to create a "rectangular" table useful for modeling from flight data. In the folder `02-fundamentals/data` there are four CSV files: `flights.csv`, `airports.csv`, `planes.csv` and `weather.csv`. Join/merge these tables such that there is one rectangular table with one row for each flight.

Put code in each of the sections provided.

1. Read Data

Using the `readr` package read the `flights` data.

```
library(readr)
library(data.table)
library(dplyr)
```

```
## ------------------------------------------------------------------------
```

```
## data.table + dplyr code now lives in dtplyr.
## Please library(dtplyr)!
```

```
## ------------------------------------------------------------------------
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:data.table':
##
##     between, first, last
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:data.table':
##
##     hour, isoweek, mday, minute, month, quarter, second, wday,
##     week, yday, year
```

```
## The following object is masked from 'package:base':
##
##     date
```

```
library(magrittr)
```

```
airports <- read_csv("~/Documents/CSX460/02-building-blocks/02-exercise-nycflights/data/airports.csv") %
```

```
## Parsed with column specification:
## cols(
##   faa = col_character(),
##   name = col_character(),
##   lat = col_double(),
##   lon = col_double(),
##   alt = col_integer(),
##   tz = col_integer(),
##   dst = col_character()
## )
```

```
flights <- read_csv("~/Documents/CSX460/02-building-blocks/02-exercise-nycflights/data/flights.csv") %>%
```

```
## Parsed with column specification:
## cols(
##   year = col_integer(),
##   month = col_integer(),
##   day = col_integer(),
##   dep_time = col_integer(),
##   sched_dep_time = col_integer(),
##   dep_delay = col_integer(),
##   arr_time = col_integer(),
##   sched_arr_time = col_integer(),
##   arr_delay = col_integer(),
##   carrier = col_character(),
##   flight = col_integer(),
##   tailnum = col_character(),
##   origin = col_character(),
##   dest = col_character(),
##   air_time = col_integer(),
##   distance = col_integer(),
##   hour = col_integer(),
##   minute = col_integer(),
##   time_hour = col_datetime(format = "")
## )
```

```
planes <- read_csv("~/Documents/CSX460/02-building-blocks/02-exercise-nycflights/data/planes.csv") %>% s
```

```
## Parsed with column specification:
## cols(
##   tailnum = col_character(),
##   year = col_integer(),
##   type = col_character(),
##   manufacturer = col_character(),
##   model = col_character(),
##   engines = col_integer(),
##   seats = col_integer(),
##   speed = col_integer(),
##   engine = col_character()
## )
```

```
weather <- read_csv("~/Documents/CSX460/02-building-blocks/02-exercise-nycflights/data/weather.csv") %>%
```

```
## Parsed with column specification:
## cols(
##   origin = col_character(),
##   year = col_integer(),
##   month = col_integer(),
##   day = col_integer(),
##   hour = col_integer(),
##   temp = col_double(),
##   dewp = col_double(),
##   humid = col_double(),
##   wind_dir = col_integer(),
##   wind_speed = col_double(),
##   wind_gust = col_double(),
##   precip = col_double(),
##   pressure = col_double(),
##   visib = col_double(),
##   time_hour = col_datetime(format = "")
## )
```

```
head(airports)
```

```
##    faa                            name     lat      lon  alt tz dst
## 1: 04G             Lansdowne Airport 41.1305 -80.6196 1044 -5   A
## 2: 06A  Moton Field Municipal Airport 32.4606 -85.6800  264 -5   A
## 3: 06C           Schaumburg Regional 41.9893 -88.1012  801 -6   A
## 4: 06N               Randall Airport 41.4319 -74.3916  523 -5   A
## 5: 09J          Jekyll Island Airport 31.0745 -81.4278   11 -4   A
## 6: 0A9 Elizabethton Municipal Airport 36.3712 -82.1734 1593 -4   A
```

```
head(weather)
```

```
##    origin year month day hour  temp  dewp humid wind_dir wind_speed
## 1:    EWR 2013     1   1    0 37.04 21.92 53.97      230    10.3570
## 2:    EWR 2013     1   1    1 37.04 21.92 53.97      230    13.8094
## 3:    EWR 2013     1   1    2 37.94 21.92 52.09      230    12.6586
## 4:    EWR 2013     1   1    3 37.94 23.00 54.51      230    13.8094
## 5:    EWR 2013     1   1    4 37.94 24.08 57.04      240    14.9601
## 6:    EWR 2013     1   1    6 39.02 26.06 59.37      270    10.3570
##    wind_gust precip pressure visib           time_hour
## 1:   11.9187      0   1013.9    10 2012-12-31 16:00:00
## 2:   15.8915      0   1013.0    10 2012-12-31 17:00:00
## 3:   14.5672      0   1012.6    10 2012-12-31 18:00:00
## 4:   15.8915      0   1012.7    10 2012-12-31 19:00:00
## 5:   17.2158      0   1012.8    10 2012-12-31 20:00:00
## 6:   11.9187      0   1012.0    10 2012-12-31 22:00:00
```

```
head(flights)
```

```
##    year month day dep_time sched_dep_time dep_delay arr_time
## 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
##    sched_arr_time arr_delay carrier flight tailnum origin dest air_time
## 1:            819        11      UA   1545  N14228    EWR  IAH      227
## 2:            830        20      UA   1714  N24211    LGA  IAH      227
## 3:            850        33      AA   1141  N619AA    JFK  MIA      160
## 4:           1022       -18      B6    725  N804JB    JFK  BQN      183
## 5:            837       -25      DL    461  N668DN    LGA  ATL      116
## 6:            728        12      UA   1696  N39463    EWR  ORD      150
##    distance hour minute           time_hour
## 1:     1400    5     15 2013-01-01 05:00:00
## 2:     1416    5     29 2013-01-01 05:00:00
## 3:     1089    5     40 2013-01-01 05:00:00
## 4:     1576    5     45 2013-01-01 05:00:00
## 5:      762    6      0 2013-01-01 06:00:00
## 6:      719    5     58 2013-01-01 05:00:00
```

```
head(planes)
```

```
##    tailnum year                    type     manufacturer     model engines
## 1:  N10156 2004 Fixed wing multi engine          EMBRAER EMB-145XR       2
## 2:  N102UW 1998 Fixed wing multi engine AIRBUS INDUSTRIE  A320-214       2
## 3:  N103US 1999 Fixed wing multi engine AIRBUS INDUSTRIE  A320-214       2
## 4:  N104UW 1999 Fixed wing multi engine AIRBUS INDUSTRIE  A320-214       2
## 5:  N10575 2002 Fixed wing multi engine          EMBRAER EMB-145LR       2
## 6:  N105UW 1999 Fixed wing multi engine AIRBUS INDUSTRIE  A320-214       2
##    seats speed    engine
## 1:    55    NA Turbo-fan
## 2:   182    NA Turbo-fan
## 3:   182    NA Turbo-fan
## 4:   182    NA Turbo-fan
## 5:    55    NA Turbo-fan
## 6:   182    NA Turbo-fan
```

Numeric Variables

Plot a histogram of arrival delays and departure delays

“‘{r} clean up NA & plot histograms ##removing NA from data set dep_delay <- na.omit(flights$dep_delay)

hist(dep_delay,breaks=100,xlim=c(-50,500),xlab=“departure delay”,main=“histogram of departure delay”) ## to see the impact of the long tail hist(dep_delay,breaks=100,xlim=c(-50,500),xlab=“departure delay”,main=“histogram of departure delay”,ylim=c(0,200))

removing NA from data set

arr_delay <- na.omit(flights$arr_delay)

hist(arr_delay,breaks=100,xlim=c(-100,500),xlab=“arrival delay”,main=“histogram of arrival delay”) ## to see the impact of the long tail hist(arr_delay,breaks=100,xlab=“arrival delay”,main=“histogram of arrival delay”,ylim=c(0,400))

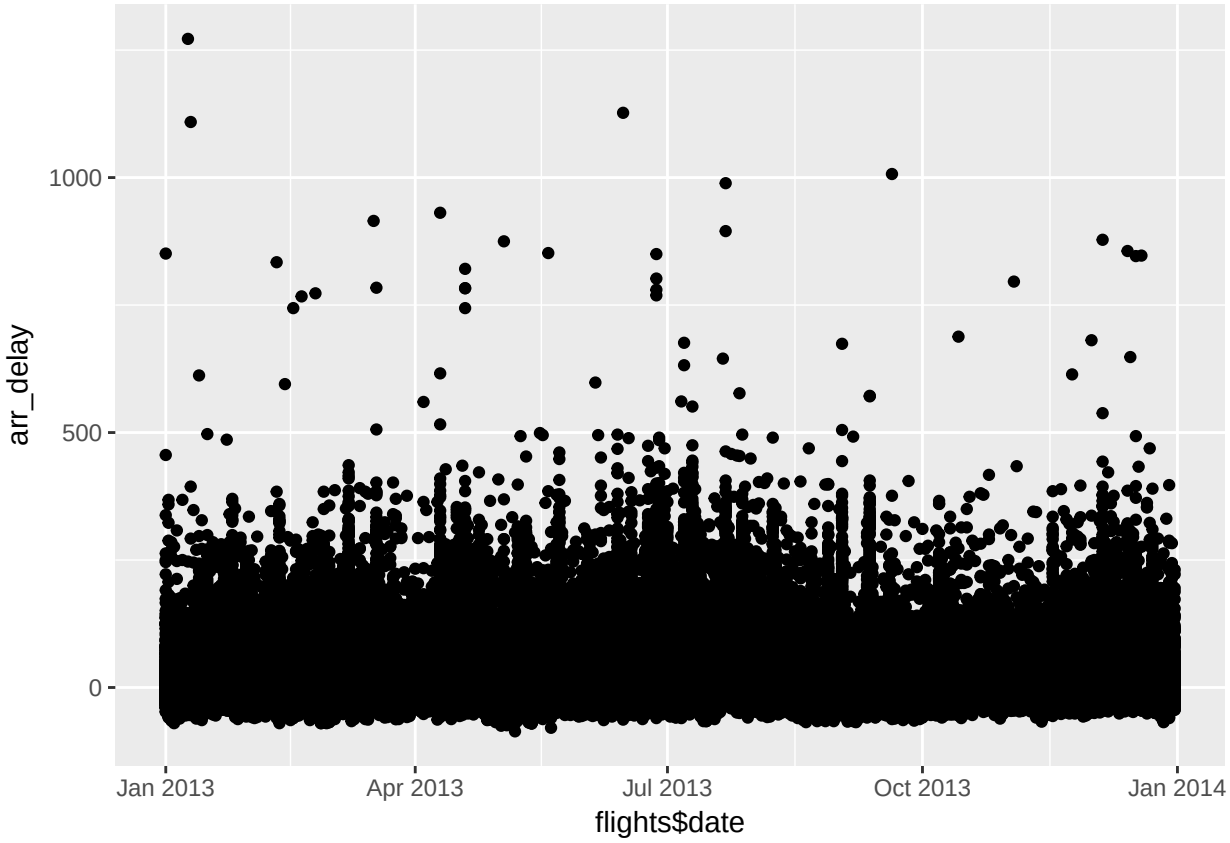
“‘

Scatterplot

Plot a scatterplot of date vs arrival delay

```
flights$date <- as.Date(with(flights, paste(year, month, day,sep="-")), "%Y-%m-%d")
ggplot(flights, aes(flights$date, arr_delay)) +geom_point()
```

```
## Warning: Removed 9430 rows containing missing values (geom_point).
```



Create tables for Categorical Variables

Create a `table`, counting the flights by origin airports and carrier.

“‘{r} counting flights by origin & carrier

flight_origin <- flights %>% group_by(origin, carrier) %>% summarize(total.count=n()) head(flight_origin)

“‘

Join Data

Read in the other data sets. Use methods to join the data together to create a single table containing one record per row. (HINT: see `?data.table` or `?dplyr::join`)

```
data_join <- flights %>%
  transform(dephour = floor(dep_time / 100)) %>%
  inner_join(.,planes, by=c("tailnum" = "tailnum")) %>%
  inner_join(.,airports,by=c("origin"="faa"))  %>%
```

```
  inner_join(.,airports,by=c("dest"="faa")) %>%
  inner_join(.,weather, by=c("origin"="origin","dephour"="hour","year.x"="year","month"="month","day"="
head(data_join)
```

```
##   year.x month day dep_time sched_dep_time dep_delay arr_time
## 1   2013     1   1      600            600         0      851
## 2   2013     1   1      601            600         1      844
## 3   2013     1   1      602            610        -8      812
## 4   2013     1   1      606            610        -4      858
## 5   2013     1   1      606            610        -4      837
## 6   2013     1   1      607            607         0      858
##   sched_arr_time arr_delay carrier flight tailnum origin dest air_time
## 1            858        -7      B6    371  N595JB    LGA  FLL      152
## 2            850        -6      B6    343  N644JB    EWR  PBI      147
## 3            820        -8      DL   1919  N971DL    LGA  MSP      170
## 4            910       -12      AA   1895  N633AA    EWR  MIA      152
## 5            845        -8      DL   1743  N3739P    JFK  ATL      128
## 6            915       -17      UA   1077  N53442    EWR  MIA      157
##   distance hour minute         time_hour.x       date dephour year.y
## 1     1076    6      0 2013-01-01 06:00:00 2013-01-01       6   2004
## 2     1023    6      0 2013-01-01 06:00:00 2013-01-01       6   2006
## 3     1020    6     10 2013-01-01 06:00:00 2013-01-01       6   1991
## 4     1085    6     10 2013-01-01 06:00:00 2013-01-01       6   1990
## 5      760    6     10 2013-01-01 06:00:00 2013-01-01       6   2000
## 6     1085    6      7 2013-01-01 06:00:00 2013-01-01       6   2009
##                       type                  manufacturer     model engines
## 1 Fixed wing multi engine                        AIRBUS  A320-232       2
## 2 Fixed wing multi engine                        AIRBUS  A320-232       2
## 3 Fixed wing multi engine MCDONNELL DOUGLAS AIRCRAFT CO     MD-88       2
## 4 Fixed wing multi engine                        BOEING   757-223       2
## 5 Fixed wing multi engine                        BOEING   737-832       2
## 6 Fixed wing multi engine                        BOEING 737-924ER       2
##   seats speed    engine              name.x   lat.x    lon.x alt.x tz.x
## 1   200    NA Turbo-fan          La Guardia 40.7772 -73.8726    22   -5
## 2   200    NA Turbo-fan Newark Liberty Intl 40.6925 -74.1687    18   -5
## 3   142    NA Turbo-fan          La Guardia 40.7772 -73.8726    22   -5
## 4   178    NA Turbo-fan Newark Liberty Intl 40.6925 -74.1687    18   -5
## 5   189    NA Turbo-jet John F Kennedy Intl 40.6398 -73.7789    13   -5
## 6   191    NA Turbo-fan Newark Liberty Intl 40.6925 -74.1687    18   -5
##   dst.x                          name.y   lat.y    lon.y alt.y tz.y dst.y
## 1     A  Fort Lauderdale Hollywood Intl 26.0726 -80.1527     9   -5     A
## 2     A                 Palm Beach Intl 26.6832 -80.0956    19   -5     A
## 3     A        Minneapolis St Paul Intl 44.8820 -93.2218   841   -6     A
## 4     A                    Miami Intl 25.7933 -80.2906     8   -5     A
## 5     A Hartsfield Jackson Atlanta Intl 33.6367 -84.4281  1026   -5     A
## 6     A                    Miami Intl 25.7933 -80.2906     8   -5     A
##    temp  dewp humid wind_dir wind_speed wind_gust precip pressure visib
## 1 39.92 26.06 57.33      260    13.8094   15.8915      0   1011.9    10
## 2 39.02 26.06 59.37      270    10.3570   11.9187      0   1012.0    10
## 3 39.92 26.06 57.33      260    13.8094   15.8915      0   1011.9    10
## 4 39.02 26.06 59.37      270    10.3570   11.9187      0   1012.0    10
## 5 39.02 26.06 59.37      260    12.6586   14.5672      0   1012.6    10
## 6 39.02 26.06 59.37      270    10.3570   11.9187      0   1012.0    10
##           time_hour.y
```

```
## 1 2012-12-31 22:00:00
## 2 2012-12-31 22:00:00
## 3 2012-12-31 22:00:00
## 4 2012-12-31 22:00:00
## 5 2012-12-31 22:00:00
## 6 2012-12-31 22:00:00
```