# Stats 101: Better flight experiences with data (airline delays in New York City)

*Nicholas J. Horton (Amherst College) and Ben Baumer (Smith College)*

*December 7, 2015*

Statistics students (and instructors) need experience wrestling with large, messy, complex, challenging data sets, for which there is no obvious goal or specially-curated statistical method. In this example, we consider a case study from a subset of the 180 million record Airline Delays dataset (see http://stat-computing.org/dataexpo/2009) that includes n=336,776 domestic commercial flights originating in New York City area airports (Newark, JFK, and LaGuardia) in 2013. These data are made available as a series of comma separated variable (CSV) files or through Hadley Wickham's `nycflights13` package on CRAN and allow students to explore a variety of statistical questions related to flights from NYC airports.

These five separate datasets can easily be merged (see the appendix for a list of the first few observations in each of these tables.) More details and extended examples can be found at http://www.amherst.edu/~nhorton/precursors.

```
require(mosaic); require(nycflights13)
```

```
# derive variables of interest...
len <- nchar(flights$dep_time)
hour <- as.numeric(substring(flights$dep_time, 1, len-2))
min <- as.numeric(substring(flights$dep_time, len-1, len))
flights <- mutate(flights, deptime = hour+min/60)
flights <- mutate(flights, realdelay = ifelse(is.na(arr_delay), 240, arr_delay))
```

Students can use this dataset to address questions that they find real and relevant. (It is not hard to find motivation for investigating patterns of flight delays. Ask students: have you ever been stuck in an airport because your flight was delayed or cancelled and wondered if you could have predicted the delay if you'd had more data?) This dataset is attractive because it is more similar to what analysts actually see in the wild than what is typically taught in the introductory statistics classroom.

**Flights to San Francisco Bay** We start with an analysis focused on three airports in the San Francisco Bay area (OAK, SFO, and SJC) for flights that depart from New York City airports.

```
require(xtable)
options(xtable.comment = FALSE)
xtab <- xtable(filter(airports, faa %in% c('SFO', 'OAK', 'SJC')))
print(xtab)
```

| | faa | name | lat | lon | alt | tz | dst |
|---|---|---|---|---|---|---|---|
| 1 | OAK | Metropolitan Oakland Intl | 37.72 | -122.22 | 9 | -8.00 | A |
| 2 | SFO | San Francisco Intl | 37.62 | -122.37 | 13 | -8.00 | A |
| 3 | SJC | Norman Y Mineta San Jose Intl | 37.36 | -121.93 | 62 | -8.00 | A |

How many flights are there to each airport in January, 2013?

```
airportcounts <- flights %>%
   filter(dest %in% c('SFO', 'OAK', 'SJC')) %>%
   group_by(year, month, dest) %>%
   summarise(count = n())
xtable(filter(airportcounts, month==1))
```

| | year | month | dest | count |
|---|---|---|---|---|
| 1 | 2013 | 1 | OAK | 20 |
| 2 | 2013 | 1 | SFO | 889 |
| 3 | 2013 | 1 | SJC | 20 |

Almost all are to San Francisco International (SFO). Let's take a closer look at what carriers service this route.

```
airlines <- mutate(airlines, name=as.character(name), carrier=as.character(carrier))
sfoflights <- inner_join(filter(flights, dest=="SFO"), airlines)
tab1 <- tally(~ name, margins=TRUE, data=sfoflights)
tab2 <- tally(~ name, format="percent", margins=TRUE, data=sfoflights)
tab <- cbind(count=tab1, percent=tab2)
xtable(tab)
```

| | count | percent |
|---|---|---|
| American Airlines Inc. | 1422.00 | 10.67 |
| Delta Air Lines Inc. | 1858.00 | 13.94 |
| JetBlue Airways | 1035.00 | 7.76 |
| United Air Lines Inc. | 6819.00 | 51.15 |
| Virgin America | 2197.00 | 16.48 |
| Total | 13331.00 | 100.00 |

United is the largest carrier (it accounts for more than half of the flights).

**Are there different delays by carrier?** Each of the carriers has at least a thousand flights, so it's likely that estimates of arrival delays may be reasonable to estimate. Let's calculate summary statistics of the arrival delay for the flights to SFO by carrier.

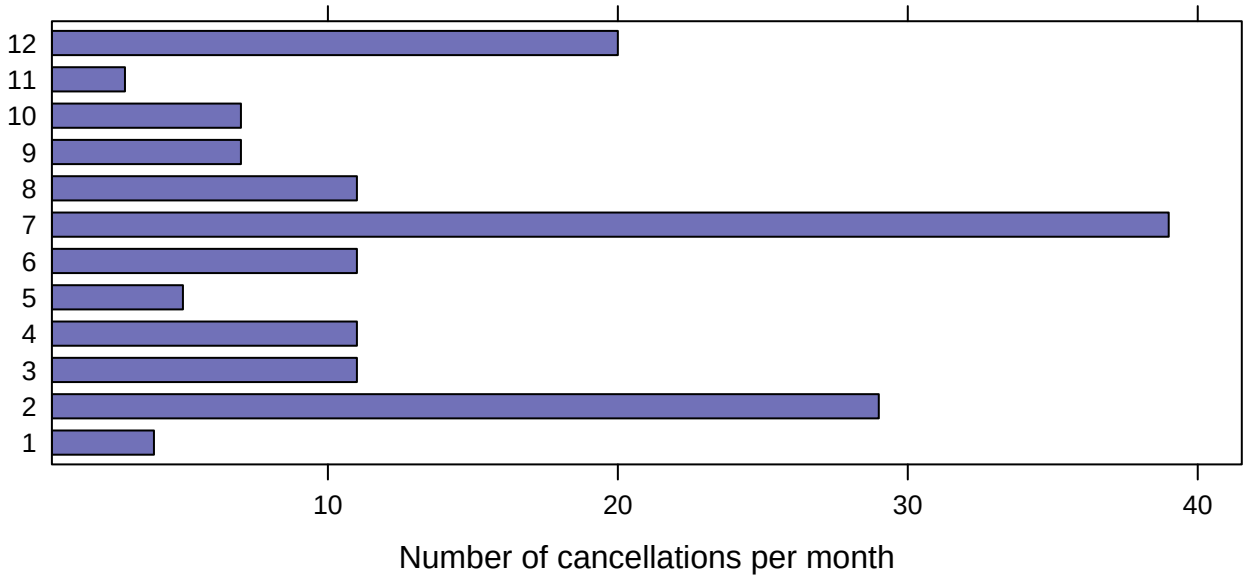
```
xtable(favstats(arr_delay ~ name, data=sfoflights))
```

| | name | min | Q1 | median | Q3 | max | mean | sd | n | missing |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | American Airlines Inc. | -63.00 | -21.00 | -4.00 | 18.00 | 1007.00 | 9.68 | 58.72 | 1398 | 24 |
| 2 | Delta Air Lines Inc. | -69.00 | -28.00 | -13.00 | 3.25 | 561.00 | -5.88 | 39.89 | 1848 | 10 |
| 3 | JetBlue Airways | -64.00 | -23.00 | -7.00 | 14.00 | 445.00 | 3.58 | 46.14 | 1020 | 15 |
| 4 | United Air Lines Inc. | -73.00 | -21.00 | -6.00 | 13.00 | 422.00 | 3.14 | 41.99 | 6728 | 91 |
| 5 | Virgin America | -86.00 | -26.00 | -12.00 | 7.00 | 676.00 | 3.58 | 60.38 | 2179 | 18 |

The "average" results (as provided by the median) is that flights arrive a few minutes early for each of these carriers. And even the 3rd quartile or the mean are relatively modest delays (all less than 20 minutes after the scheduled arrival time). But the maximum delays can be large (e.g., more than 10 hours for Virgin America and American Airlines).

We also observe that a number of flights are missing their arrival delay. Those missing values are likely cancelled flights. We might be interested in which month they occurred?

```
tab <- t(tally(month ~ 1, margin=TRUE, data=filter(sfoflights, is.na(arr_delay))))
tab <- tab[,-13]
barchart(tab, stack=FALSE, xlab="Number of cancellations per month")
```




Cancelled flights seem to be most common in July, February, and December.

How should the cancelled flights be handled? (Note that there were excluded from the calculation of the summary statistics displayed earlier.)

One option might be to recode these as 4 hour (240 minute) delays, since it's likely that if a flight is cancelled the `expected` delay might be of that duration on average. (This is an arbitrary choice: students might be asked what other options are reasonable. More sophisticated approaches could implement a "hurdle" method with a model for the probability of not being cancelled along with a model for the "average" delay for those flights that were not cancelled.)

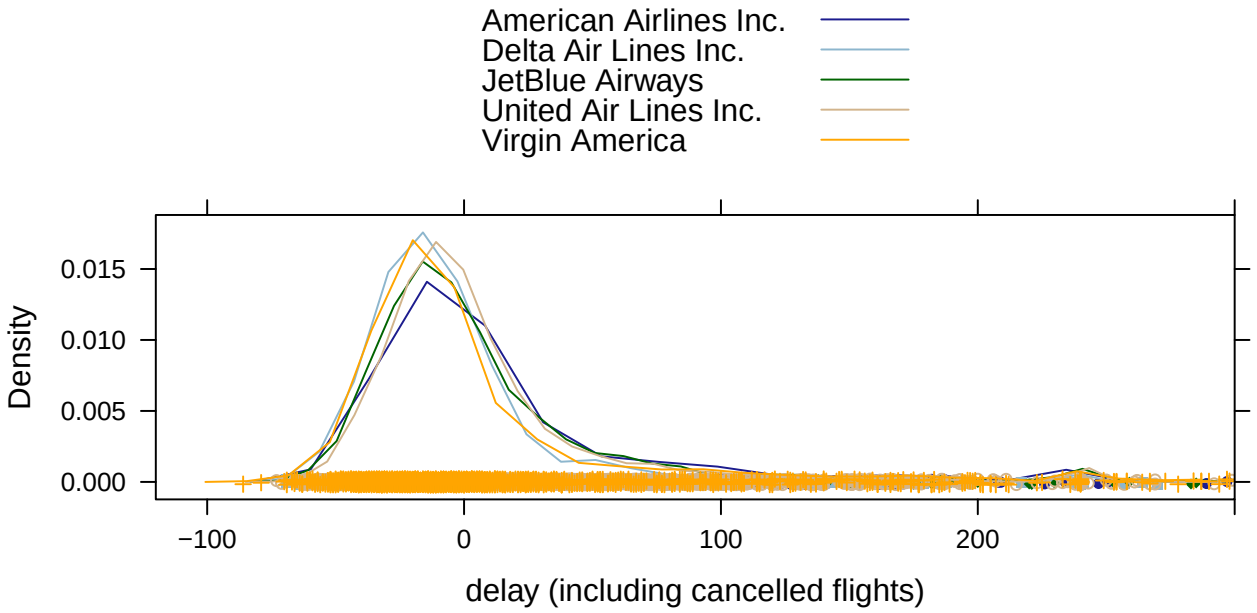
Let's revisit the distribution of real delays (accounting for cancellations) by carrier.

```
xtable(favstats(realdelay ~ name, data=sfoflights))
```

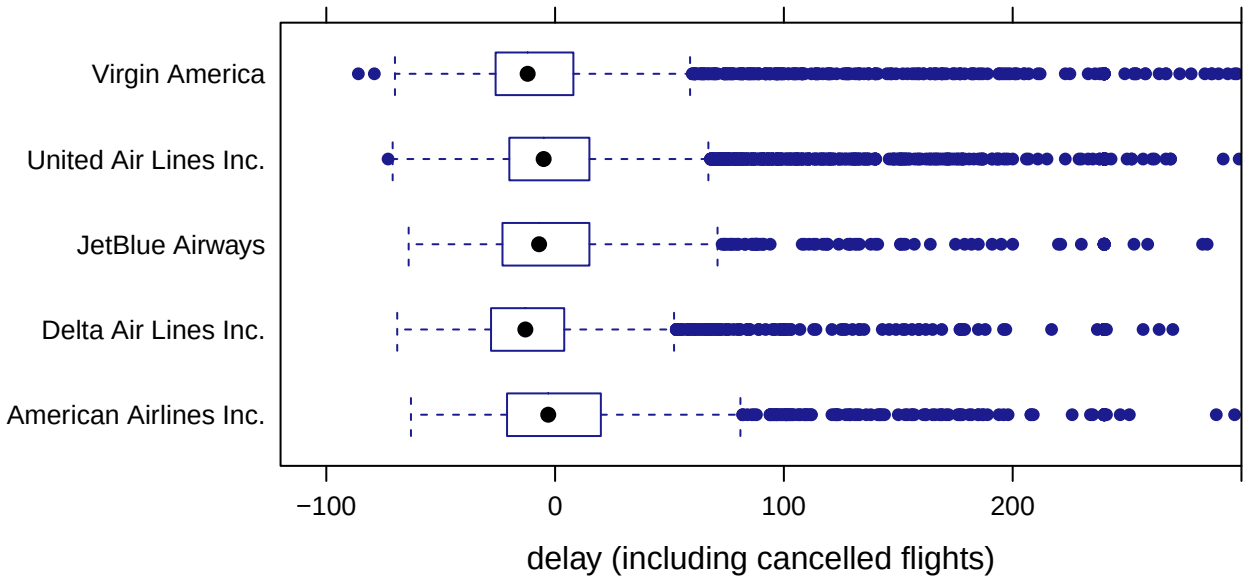
| | name | min | Q1 | median | Q3 | max | mean | sd | n | missing |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | American Airlines Inc. | -63.00 | -21.00 | -3.00 | 20.00 | 1007.00 | 13.57 | 65.35 | 1422 | 0 |
| 2 | Delta Air Lines Inc. | -69.00 | -27.75 | -13.00 | 4.00 | 561.00 | -4.56 | 43.67 | 1858 | 0 |
| 3 | JetBlue Airways | -64.00 | -23.00 | -7.00 | 15.00 | 445.00 | 7.01 | 53.82 | 1035 | 0 |
| 4 | United Air Lines Inc. | -73.00 | -20.00 | -5.00 | 15.00 | 422.00 | 6.30 | 49.78 | 6819 | 0 |
| 5 | Virgin America | -86.00 | -26.00 | -12.00 | 8.00 | 676.00 | 5.51 | 63.80 | 2197 | 0 |

A parallel graphical description of the flights delays to San Francisco airport can be used to judge the airlines.

```
densityplot(~ realdelay, group=name, auto.key=TRUE, xlim=c(-120, 300), xlab="delay (including cancelled
```




```
bwplot(name ~ realdelay, xlim=c(-120, 300), xlab="delay (including cancelled flights)",
       data=sfoflights)
```

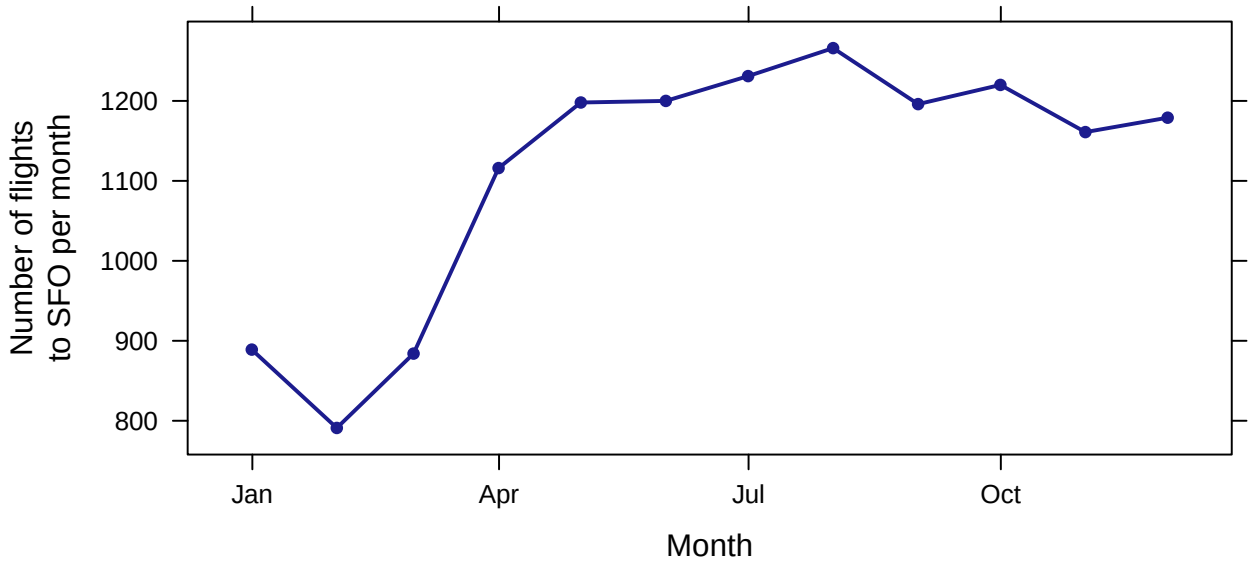



Note that the distributions have been rescaled so that only those flights between 2 hours early and 5 hours late are displayed (this excludes some of the extreme outliers).

The distributions appear to be somewhat symmetrically distributed around zero delays but with extremely long right tails. Different information is conveyed in the two representations: the overlapping density plots provide a clear sense of the shape of the distributions but are somewhat crowded. The boxplots make it easy to compare airline reliability, and to see the quantiles.

**Is there seasonality to the number of flights?** We can consider whether the number of flights changes month by month.

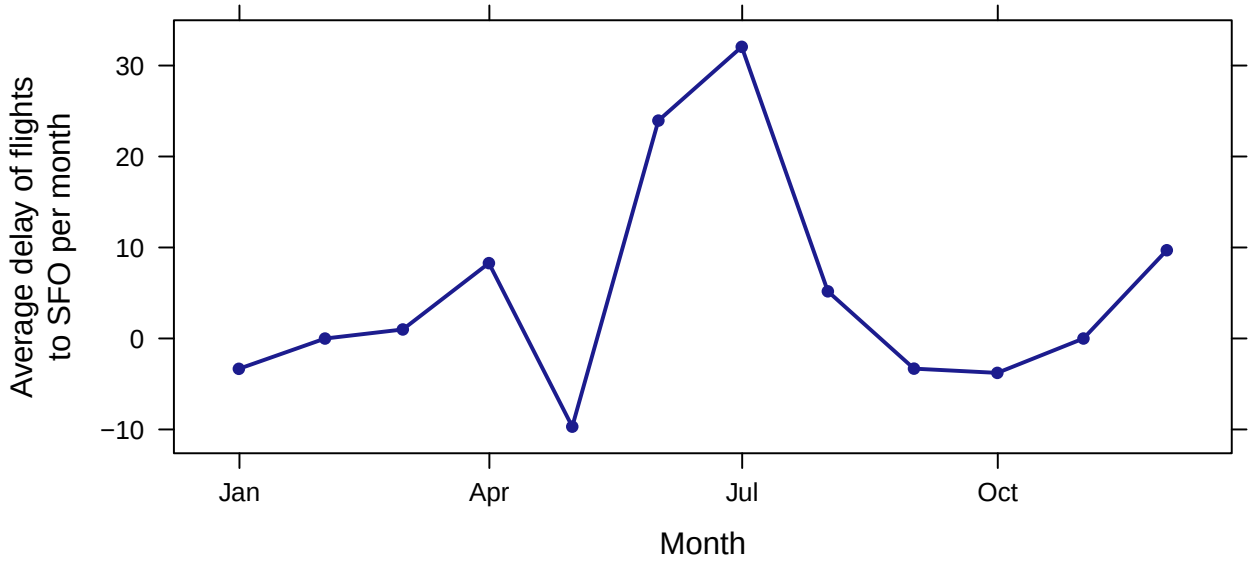
```
sfocounts <- filter(airportcounts, dest=="SFO") %>%
  mutate(Date = ymd(paste(year, "-", month, "-01", sep="")))
xyplot(count ~ Date, type=c("p","l"), lwd=2,
       xlab="Month",
       ylab="Number of flights\nto SFO per month", data=sfocounts)
```




We observe that there are some interesting patterns over the course of the year for SFO: the number of flights in January, February, and March is considerably less than the other nine months.

**Predictors of delays** How is month of flight associated with delays?

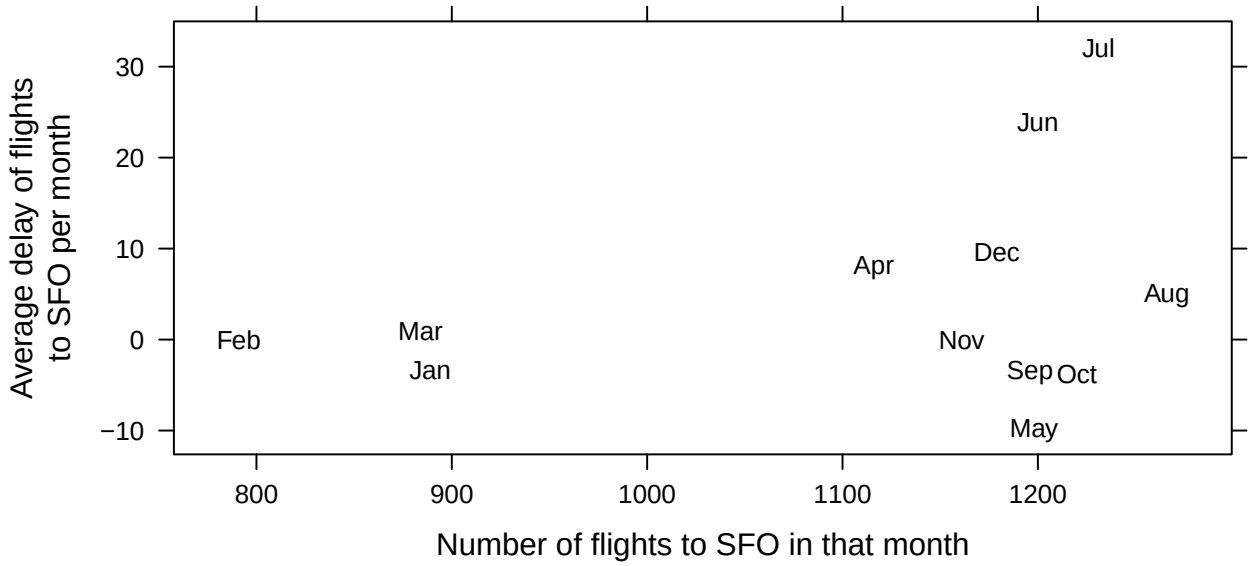
```
sfocounts <- sfoflights %>%
  mutate(Date = ymd(paste(year, "-", month, "-01", sep=""))) %>%
  group_by(Date) %>%
  summarise(count = n(), avgdelay = mean(realdelay))
xyplot(avgdelay ~ Date, type=c("p","l"), lwd=2,
       xlab="Month",
       ylab="Average delay of flights\nto SFO per month", data=sfocounts)
```




We see that the largest average delays occur in the summer (with both June and July having an average above 20 minutes).

Is there an association between the number of flights in a month and the average delay?

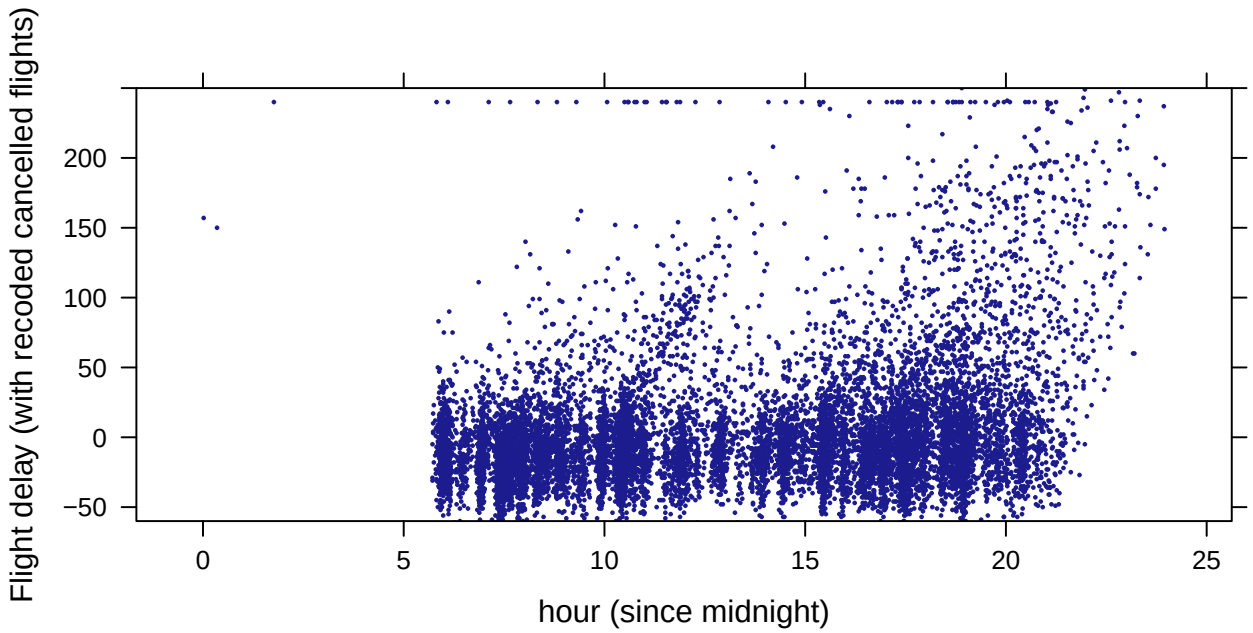
```
panel.labels <- function(x, y, labels='x',...) {
  panel.text(x, y, labels, cex=0.8, ...)
}
xyplot(avgdelay ~ count, type=c("p"), lwd=2,
       panel=panel.labels, labels=month(sfocounts$Date, label=TRUE),
        xlab="Number of flights to SFO in that month",
       ylab="Average delay of flights\nto SFO per month", data=sfocounts)
```




There is not much of a pattern, but the delays seem to be more variable on months with more flights.

Another question that travelers might consider is whether the departure time matter as a predictor of flight delays?

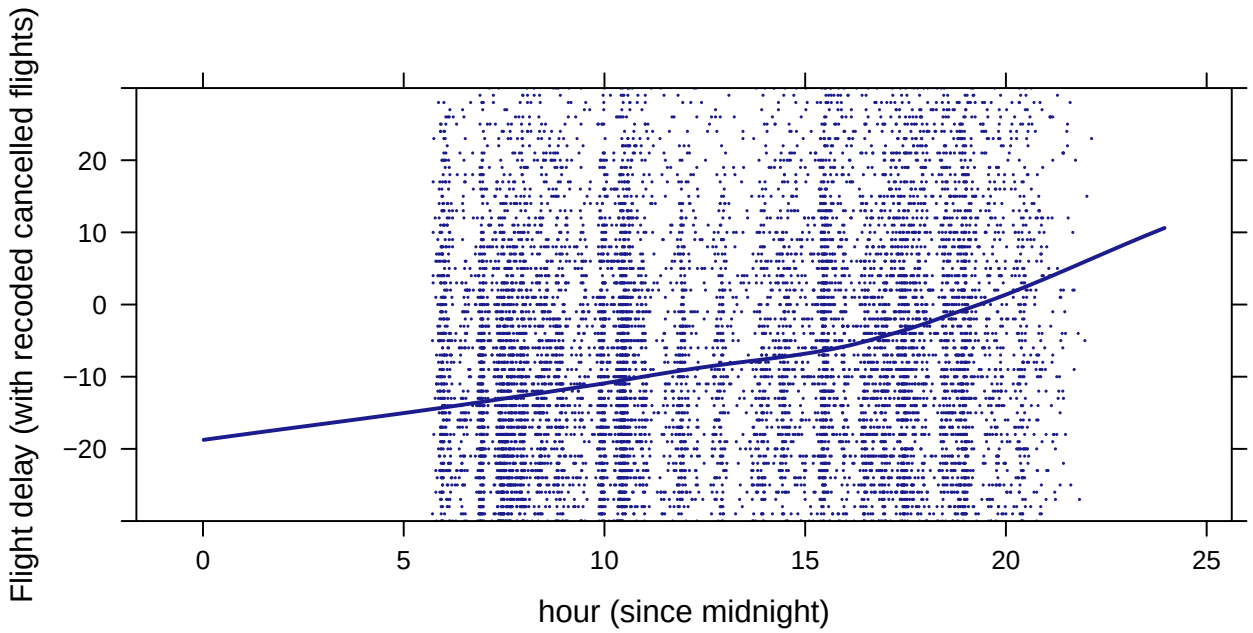
```
xyplot(realdelay ~ deptime, cex=0.3, type=c("p"), xlab="hour (since midnight)",
       ylab="Flight delay (with recoded cancelled flights)", ylim=c(-60, 250), data=sfoflights)
```




A number of observations can be made from this graphical display. Very few flights depart between midnight and 5:30am. Most flights are on time, but there does appear to be a pattern that more delays occur for flights that are scheduled to depart later in the day.

We can improve the display by zooming in and adding a scatterplot smooth.

```
xyplot(realdelay ~ deptime, cex=0.2, type=c("p", "smooth"), xlab="hour (since midnight)",
       ylab="Flight delay (with recoded cancelled flights)", ylim=c(-30, 30), data=sfoflights)
```



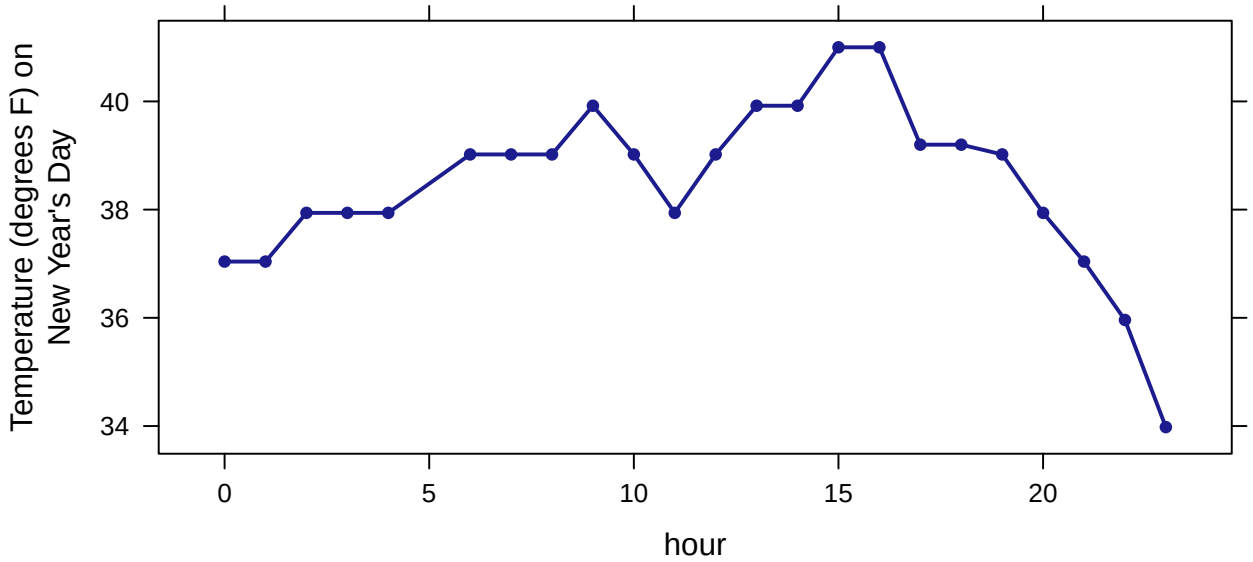

While there is some indication that delays tend to be more common (and slightly longer) as the day proceeds, the effect is modest for flights to San Francisco.

**Weather** Other factors affect airline delays. This might include the weather.
The `nycflights13` package in R includes other data scraped from the Internet (in this case detailed weather information). We can display the temperature (in degrees Fahrenheit) on New Year's Day, 2013.

```
xtable(select(weather, hour, hour, dewp, humid, wind_speed, wind_dir, precip, pressure) %>% head())
```

| | month | day | hour | dewp | humid | wind_speed | wind_dir | precip | pressure |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1.00 | 1 | 0 | 21.92 | 53.97 | 10.36 | 230.00 | 0.00 | 1013.90 |
| 2 | 1.00 | 1 | 1 | 21.92 | 53.97 | 13.81 | 230.00 | 0.00 | 1013.00 |
| 3 | 1.00 | 1 | 2 | 21.92 | 52.09 | 12.66 | 230.00 | 0.00 | 1012.60 |
| 4 | 1.00 | 1 | 3 | 23.00 | 54.51 | 13.81 | 230.00 | 0.00 | 1012.70 |
| 5 | 1.00 | 1 | 4 | 24.08 | 57.04 | 14.96 | 240.00 | 0.00 | 1012.80 |
| 6 | 1.00 | 1 | 6 | 26.06 | 59.37 | 10.36 | 270.00 | 0.00 | 1012.00 |

```
xyplot(temp ~ hour, type=c("p", "l"),
       ylab="Temperature (degrees F) on\nNew Year's Day", data=filter(weather, month==1 & day==1))
```




Let's take a look at daily averages for delays as well as total precipation and maximum wind speed. First we undertake the merge and display a set of values.

```
avgdelay <- flights %>%
  group_by(month, day) %>%
  filter(month < 13) %>%
  summarise(avgdelay = mean(realdelay, na.rm=TRUE))
precip <- weather %>%
  group_by(month, day) %>%
  filter(month < 13) %>%
  summarise(totprecip = sum(precip), maxwind = max(wind_speed))
precip <- mutate(precip, anyprecip = ifelse(totprecip==0, "No", "Yes"))
merged <- left_join(avgdelay, precip, by=c("day", "month"))
xtable(head(merged))
```

A dramatic outlier is immediately spotted: windspeeds of 1000 mph are not common! This must be an error.

| | month | day | avgdelay | totprecip | maxwind | anyprecip |
|---|---|---|---|---|---|---|
| 1 | 1.00 | 1 | 15.62 | 0.00 | 16.11 | No |
| 2 | 1.00 | 2 | 16.31 | 0.00 | 18.41 | No |
| 3 | 1.00 | 3 | 9.32 | 0.00 | 11.51 | No |
| 4 | 1.00 | 4 | -0.08 | 0.00 | 24.17 | No |
| 5 | 1.00 | 5 | -0.52 | 0.00 | 18.41 | No |
| 6 | 1.00 | 6 | 5.09 | 0.00 | 14.96 | No |

```
xtable(favstats(~ maxwind, data=merged))
```

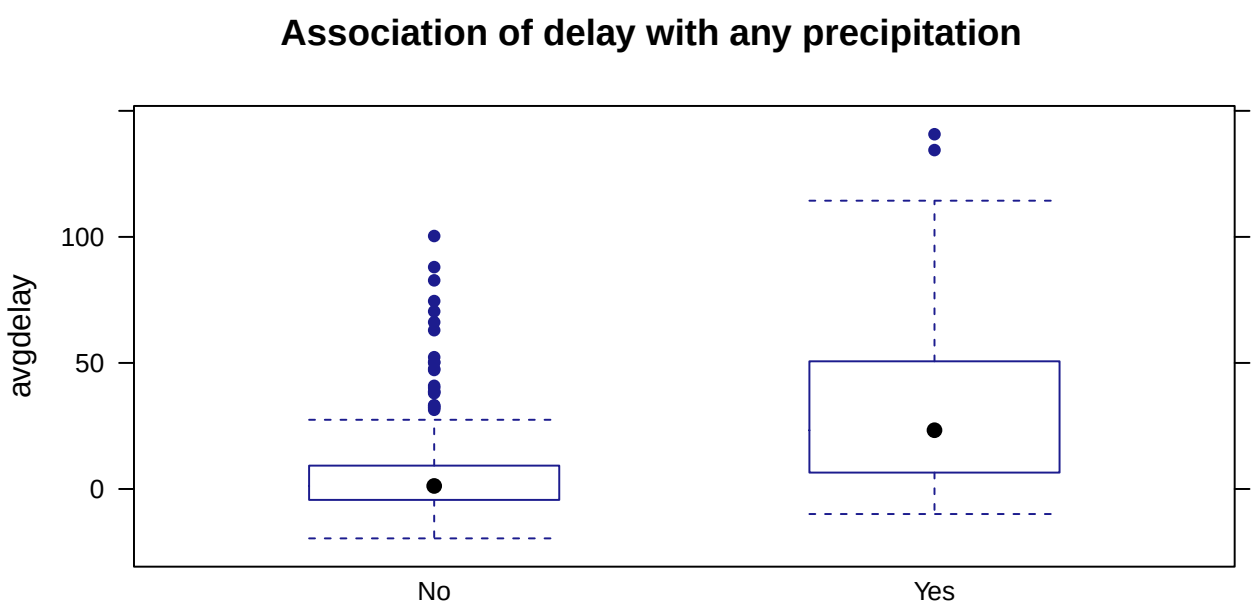
| min | Q1 | median | Q3 | max | mean | sd | n | missing |
|---|---|---|---|---|---|---|---|---|
| 5.75 | 12.66 | 16.11 | 19.56 | 1048.36 | 19.26 | 54.43 | 363 | 2 |

```
xtable(filter(merged, maxwind > 1000))
```

| | month | day | avgdelay | totprecip | maxwind | anyprecip |
|---|---|---|---|---|---|---|
| 1 | 2.00 | 12 | 0.30 | 0.00 | 1048.36 | No |

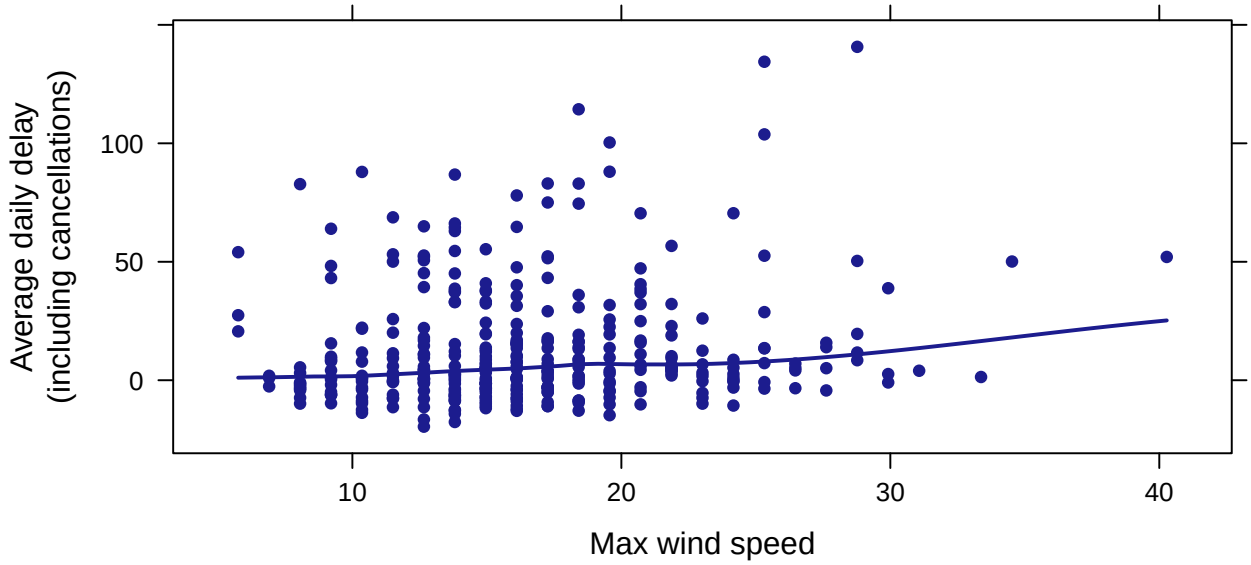
Let's remove this outlier and consider the association between any precipiation and average delays.

```
merged <- filter(merged, maxwind < 1000)
bwplot(avgdelay ~ anyprecip, main="Association of delay with any precipitation", data=merged)
```



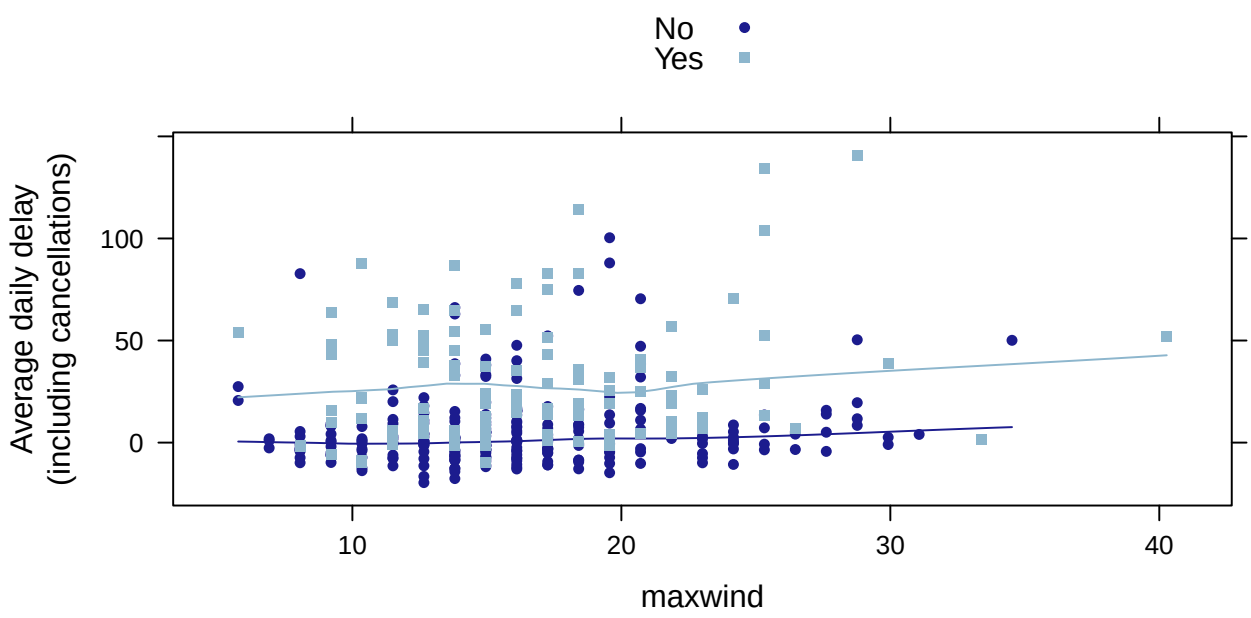

Precipitation seems to be associated with delays:

```
xyplot(avgdelay ~ maxwind, type=c("p", "smooth"), xlab="Max wind speed", ylab="Average daily delay\n(in
```




Max windspeed also seems to be associated with delays.

```
xyplot(avgdelay ~ maxwind, groups=anyprecip, auto.key=TRUE, type=c("p", "smooth"), ylab="Average daily
```




After stratifying by precipitation status, we see that windspeed does not appear to be a major determinant of delays. Precipitation seems to be the issue.

## Closing thoughts and further resources

The dataset (as a series of comma separated variable files), copies of the R Markdown and formatted files for these analyses (to allow replication of the analyses) along with further background on the Airline Delays dataset can be found at http://www.amherst.edu/~nhorton/precursors.

Horton, N.J., Baumer, B.S., and Wichham H. (2015) Setting the stage for data science: integration of data management skills in introductory and second courses in statistics“, *CHANCE*, 28(2):40-50, http://chance.amstat.org/2015/04/setting-the-stage.

Kane, M. Strategies for analyzing a 12-gigabyte data set: airline flight delays (2015) in *Data Science in R: A Case Studies Approach to Computational Reasoning and Problem Solving*, Nolan D. and Temple Lang D, CRC Press.

Wickham, H. (2011). ASA 2009 Data Expo, *Journal of Computational and Graphical Statistics*, 20(2):281-283.

**Appendix** In this appendix, the first few rows of each of the datasets is displayed.

```
airlines
```

```
## Source: local data frame [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.
```

```
airports
```

```
## Source: local data frame [1,397 x 7]
##
##      faa                            name   lat    lon   alt    tz   dst
##    (chr)                           (chr) (dbl)  (dbl) (int) (dbl) (chr)
## 1    04G              Lansdowne Airport  41.1  -80.6  1044    -5     A
## 2    06A  Moton Field Municipal Airport  32.5  -85.7   264    -5     A
## 3    06C            Schaumburg Regional  42.0  -88.1   801    -6     A
## 4    06N                Randall Airport  41.4  -74.4   523    -5     A
## 5    09J          Jekyll Island Airport  31.1  -81.4    11    -4     A
## 6    0A9 Elizabethton Municipal Airport  36.4  -82.2  1593    -4     A
## 7    0G6        Williams County Airport  41.5  -84.5   730    -5     A
## 8    0G7  Finger Lakes Regional Airport  42.9  -76.8   492    -5     A
## 9    0P2   Shoestring Aviation Airfield  39.8  -76.6  1000    -5     U
## 10   0S9          Jefferson County Intl  48.1 -122.8   108    -8     A
## ..   ...                            ...   ...    ...   ...   ...   ...
```

```
planes
```

```
## Source: local data frame [3,322 x 9]
```

```
##
##    tailnum  year                    type     manufacturer     model
##      (chr) (int)                   (chr)            (chr)     (chr)
## 1   N10156  2004 Fixed wing multi engine          EMBRAER EMB-145XR
## 2   N102UW  1998 Fixed wing multi engine AIRBUS INDUSTRIE  A320-214
## 3   N103US  1999 Fixed wing multi engine AIRBUS INDUSTRIE  A320-214
## 4   N104UW  1999 Fixed wing multi engine AIRBUS INDUSTRIE  A320-214
## 5   N10575  2002 Fixed wing multi engine          EMBRAER EMB-145LR
## 6   N105UW  1999 Fixed wing multi engine AIRBUS INDUSTRIE  A320-214
## 7   N107US  1999 Fixed wing multi engine AIRBUS INDUSTRIE  A320-214
## 8   N108UW  1999 Fixed wing multi engine AIRBUS INDUSTRIE  A320-214
## 9   N109UW  1999 Fixed wing multi engine AIRBUS INDUSTRIE  A320-214
## 10  N110UW  1999 Fixed wing multi engine AIRBUS INDUSTRIE  A320-214
## ..     ...   ...                    ...              ...       ...
## Variables not shown: engines (int), seats (int), speed (int), engine (chr)
```

```
flights
```

```
## Source: local data frame [336,776 x 18]
##
##     year month   day dep_time dep_delay arr_time arr_delay carrier tailnum
##    (int) (int) (int)    (int)     (dbl)    (int)     (dbl)   (chr)   (chr)
## 1   2013     1     1      517         2      830        11      UA  N14228
## 2   2013     1     1      533         4      850        20      UA  N24211
## 3   2013     1     1      542         2      923        33      AA  N619AA
## 4   2013     1     1      544        -1     1004       -18      B6  N804JB
## 5   2013     1     1      554        -6      812       -25      DL  N668DN
## 6   2013     1     1      554        -4      740        12      UA  N39463
## 7   2013     1     1      555        -5      913        19      B6  N516JB
## 8   2013     1     1      557        -3      709       -14      EV  N829AS
## 9   2013     1     1      557        -3      838        -8      B6  N593JB
## 10  2013     1     1      558        -2      753         8      AA  N3ALAA
## ..   ...   ...   ...      ...       ...      ...       ...     ...     ...
## Variables not shown: flight (int), origin (chr), dest (chr), air_time
##   (dbl), distance (dbl), hour (dbl), minute (dbl), deptime (dbl),
##   realdelay (dbl)
```

```
weather
```

```
## Source: local data frame [8,719 x 14]
## Groups: month, day, hour [8719]
##
##    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     0  37.0  21.9  54.0      230      10.36
## 2     EWR  2013     1     1     1  37.0  21.9  54.0      230      13.81
## 3     EWR  2013     1     1     2  37.9  21.9  52.1      230      12.66
## 4     EWR  2013     1     1     3  37.9  23.0  54.5      230      13.81
## 5     EWR  2013     1     1     4  37.9  24.1  57.0      240      14.96
## 6     EWR  2013     1     1     6  39.0  26.1  59.4      270      10.36
## 7     EWR  2013     1     1     7  39.0  27.0  61.6      250       8.06
## 8     EWR  2013     1     1     8  39.0  28.0  64.4      240      11.51
## 9     EWR  2013     1     1     9  39.9  28.0  62.2      250      12.66
```

```
## 10    EWR  2013     1     1    10  39.0  28.0  64.4      260      12.66
## ..    ...   ...   ...   ...   ...   ...   ...   ...      ...        ...
## Variables not shown: wind_gust (dbl), precip (dbl), pressure (dbl), visib
##   (dbl)
```

```
write.csv(airlines, file="stats101-airlines.csv")
write.csv(airports, file="stats101-airports.csv")
write.csv(planes,   file="stats101-planes.csv")
write.csv(flights,  file="stats101-flights.csv")
write.csv(weather,  file="stats101-weather.csv")
write.csv(sfoflights,  file="stats101-sfoflights.csv")
```