# DA607 Week 06 Assignment

*Dan Fanelli*

*March 2, 2016*

## Tidying and Transforming Data

### Filter, Scrub, Format:

```r
library(sqldf)
library(zoo)
library(knitr)
library(dplyr)

file_path <- "arrival_delays.csv"
# file_path <- "http://raw.githubusercontent.com/fandang/DA607/master/Wk05/arrival_delays.csv"
delays <- read.csv(file_path, header = TRUE, sep = ",")
colnames(delays) <- c("airline","arrival_status","LA","PHO","SD","SF","SEA")

# Show the data just after import and column renames
kable(delays)
```

| airline | arrival_status | LA | PHO | SD | SF | SEA |
|---|---|---|---|---|---|---|
| ALASKA | on time | 497 | 221 | 212 | 503 | 1,841 |
| | delayed | 62 | 12 | 20 | 102 | 305 |
| | | NA | | NA | NA | |
| AM WEST | on time | 694 | 4,840 | 383 | 320 | 201 |
| | delayed | 117 | 415 | 65 | 129 | 61 |

```r
delays <- filter(delays, !is.na(LA), !is.na(PHO), !is.na(SD), !is.na(SF), !is.na(SEA))

delays$airline[delays$airline == ""] <- NA
delays$airline <- na.locf(delays$airline)

# get rid of commas in the numbers...otherwise the next batch of "transform" calls gives back the wrong
delays$LA <- gsub(",", "", delays$LA)
delays$PHO <- gsub(",", "", delays$PHO)
delays$SD <- gsub(",", "", delays$SD)
delays$SF <- gsub(",", "", delays$SF)
delays$SEA <- gsub(",", "", delays$SEA)

# make sure the necessary columns are numeric - it doesn't err out without this, it just results in del
delays <- transform(delays, LA = as.numeric(LA))
delays <- transform(delays, PHO = as.numeric(PHO))
delays <- transform(delays, SD = as.numeric(SD))
delays <- transform(delays, SF = as.numeric(SF))
delays <- transform(delays, SEA = as.numeric(SEA))
```

```
kable(delays)
```

| airline | arrival_status | LA | PHO | SD | SF | SEA |
|---|---|---|---|---|---|---|
| ALASKA | on time | 497 | 221 | 212 | 503 | 1841 |
| ALASKA | delayed | 62 | 12 | 20 | 102 | 305 |
| AM WEST | on time | 694 | 4840 | 383 | 320 | 201 |
| AM WEST | delayed | 117 | 415 | 65 | 129 | 61 |

### Now do a few calculations:

```
# There is a good amount of repeat in the select clause, there must be a way to get the "on_time" resul
kable(sqldf("select d.airline, (select (LA+PHO+SD+SF+SEA) from delays d2 where d.airline = d2.airline ar
```

| airline | num_on_time | num_delayed | delayed_pct |
|---|---|---|---|
| ALASKA | 3274 | 501 | 0.1327152 |
| AM WEST | 6438 | 787 | 0.1089273 |

### Some R test cases to confirm the sql calculations:

```
# now confirm it straight from the csv file:
alaska_on_time <- c(497,221,212,503,1841)
alaska_delayed <- c(62,12,20,102,305)
alaska_on_time <- sum(alaska_on_time)
alaska_delayed <- sum(alaska_delayed)
alaska_delay_pct <- alaska_delayed / (alaska_on_time + alaska_delayed)
cat(alaska_delayed,"/",(alaska_on_time + alaska_delayed),"=",alaska_delay_pct)
```

```
## 501 / 3775 = 0.1327152
```

```
amwest_on_time <- c(694,4840,383,320,201)
amwest_delayed <- c(117,415,65,129,61)
amwest_on_time <- sum(amwest_on_time)
amwest_delayed <- sum(amwest_delayed)
amwest_delay_pct <- amwest_delayed / (amwest_on_time + amwest_delayed)
cat(amwest_delayed,"/",(amwest_on_time + amwest_delayed),"=",amwest_delay_pct)
```

```
## 787 / 7225 = 0.1089273
```