Transformation

Prof. Eric A. Suess

August 28, 2019

Data Transformation

Today we will begin discussing Data Wrangling.

The R package that will be using from the tidyverse is the dplyr package.

The grammar of data wrangling

The 5 verbs of data wrangling

- **filter()** Pick observations by their values
- **arrange()** Reorder the rows
- **select()** Pick variables by their names
- **mutate()** Create new variables with functions of existing variables
- **summarise()** Collapse many values down to a single summary
- **group_by()**

RStudio Cheatsheet for dplyr

The RStudio dplyr cheatsheet is very useful.

Flights data

```r
library(nycflights13)
library(tidyverse)
```

Flights data

```
flights
```

```
## # A tibble: 336,776 x 19
##     year month   day dep_time sched_dep_time dep_delay a
##    <int> <int> <int>    <int>          <int>     <dbl>
##  1  2013     1     1      517            515         2
##  2  2013     1     1      533            529         4
##  3  2013     1     1      542            540         2
##  4  2013     1     1      544            545        -1
##  5  2013     1     1      554            600        -6
##  6  2013     1     1      554            558        -4
##  7  2013     1     1      555            600        -5
##  8  2013     1     1      557            600        -3
##  9  2013     1     1      557            600        -3
## 10  2013     1     1      558            600        -2
## # ... with 336,766 more rows, and 12 more variables: sch
## #   arr_delay <dbl>, carrier <chr>, flight <int>, tailnu
## #   origin <chr>, dest <chr>, air_time <dbl>, distance <
```

filter()

```r
filter(flights, month == 1, day == 1)
```

```
## # A tibble: 842 x 19
##     year month   day dep_time sched_dep_time dep_delay a
##    <int> <int> <int>    <int>          <int>     <dbl>
##  1  2013     1     1      517            515         2
##  2  2013     1     1      533            529         4
##  3  2013     1     1      542            540         2
##  4  2013     1     1      544            545        -1
##  5  2013     1     1      554            600        -6
##  6  2013     1     1      554            558        -4
##  7  2013     1     1      555            600        -5
##  8  2013     1     1      557            600        -3
##  9  2013     1     1      557            600        -3
## 10  2013     1     1      558            600        -2
## # ... with 832 more rows, and 12 more variables: sched_a
## #   arr_delay <dbl>, carrier <chr>, flight <int>, tailnu
## #   origin <chr>, dest <chr>, air_time <dbl>, distance <
```

arrange()

```r
arrange(flights, year, month, day)
```

```
## # A tibble: 336,776 x 19
##     year month   day dep_time sched_dep_time dep_delay a
##    <int> <int> <int>    <int>          <int>     <dbl>
##  1  2013     1     1      517            515         2
##  2  2013     1     1      533            529         4
##  3  2013     1     1      542            540         2
##  4  2013     1     1      544            545        -1
##  5  2013     1     1      554            600        -6
##  6  2013     1     1      554            558        -4
##  7  2013     1     1      555            600        -5
##  8  2013     1     1      557            600        -3
##  9  2013     1     1      557            600        -3
## 10  2013     1     1      558            600        -2
## # ... with 336,766 more rows, and 12 more variables: sch
## #   arr_delay <dbl>, carrier <chr>, flight <int>, tailnu
## #   origin <chr>, dest <chr>, air_time <dbl>, distance <
```

arrange()

```r
arrange(flights, desc(dep_delay))
```

```
## # A tibble: 336,776 x 19
##     year month   day dep_time sched_dep_time dep_delay a
##    <int> <int> <int>    <int>          <int>     <dbl>
##  1  2013     1     9      641            900      1301
##  2  2013     6    15     1432           1935      1137
##  3  2013     1    10     1121           1635      1126
##  4  2013     9    20     1139           1845      1014
##  5  2013     7    22      845           1600      1005
##  6  2013     4    10     1100           1900       960
##  7  2013     3    17     2321            810       911
##  8  2013     6    27      959           1900       899
##  9  2013     7    22     2257            759       898
## 10  2013    12     5      756           1700       896
## # ... with 336,766 more rows, and 12 more variables: sch
## #   arr_delay <dbl>, carrier <chr>, flight <int>, tailnu
## #   origin <chr>, dest <chr>, air_time <dbl>, distance <
```

select()

```r
select(flights, year, month, day)
```

```
## # A tibble: 336,776 x 3
##     year month   day
##    <int> <int> <int>
##  1  2013     1     1
##  2  2013     1     1
##  3  2013     1     1
##  4  2013     1     1
##  5  2013     1     1
##  6  2013     1     1
##  7  2013     1     1
##  8  2013     1     1
##  9  2013     1     1
## 10  2013     1     1
## # ... with 336,766 more rows
```

select()

```r
select(flights, time_hour, air_time, everything())
```

```
## # A tibble: 336,776 x 19
##    time_hour           air_time  year month   day dep_ti
##    <dttm>                 <dbl> <int> <int> <int>    <in
##  1 2013-01-01 05:00:00      227  2013     1     1      5
##  2 2013-01-01 05:00:00      227  2013     1     1      5
##  3 2013-01-01 05:00:00      160  2013     1     1      5
##  4 2013-01-01 05:00:00      183  2013     1     1      5
##  5 2013-01-01 06:00:00      116  2013     1     1      5
##  6 2013-01-01 05:00:00      150  2013     1     1      5
##  7 2013-01-01 06:00:00      158  2013     1     1      5
##  8 2013-01-01 06:00:00       53  2013     1     1      5
##  9 2013-01-01 06:00:00      140  2013     1     1      5
## 10 2013-01-01 06:00:00      138  2013     1     1      5
## # ... with 336,766 more rows, and 12 more variables: dep
## #   arr_time <int>, sched_arr_time <int>, arr_delay <dbl
## #   flight <int>, tailnum <chr>, origin <chr>, dest <chr
```

mutate()

```r
flights_sml <- select(flights,
  year:day,
  ends_with("delay"),
  distance,
  air_time
)
mutate(flights_sml,
  gain = dep_delay - arr_delay,
  speed = distance / air_time * 60
)
```

```
## # A tibble: 336,776 x 9
##      year month   day dep_delay arr_delay distance air_ti
##     <int> <int> <int>     <dbl>     <dbl>    <dbl>    <db
##  1  2013     1     1         2        11     1400      2
##  2  2013     1     1         4        20     1416      2
##  3  2013     1     1         2        33     1089      1
##  4  2013     1     1        -1       -18     1576      1
```

summarize()

```r
summarise(flights, delay = mean(dep_delay, na.rm = TRUE))
```

```
## # A tibble: 1 x 1
##   delay
##   <dbl>
## 1  12.6
```

```r
by_day <- group_by(flights, year, month, day)
summarise(by_day, delay = mean(dep_delay, na.rm = TRUE))
```

```
## # A tibble: 365 x 4
## # Groups:   year, month [?]
##     year month   day delay
##    <int> <int> <int> <dbl>
##  1  2013     1     1 11.5
##  2  2013     1     2 13.9
##  3  2013     1     3 11.0
##  4  2013     1     4  8.95
```

Combining multiple operations with the pipe %>%

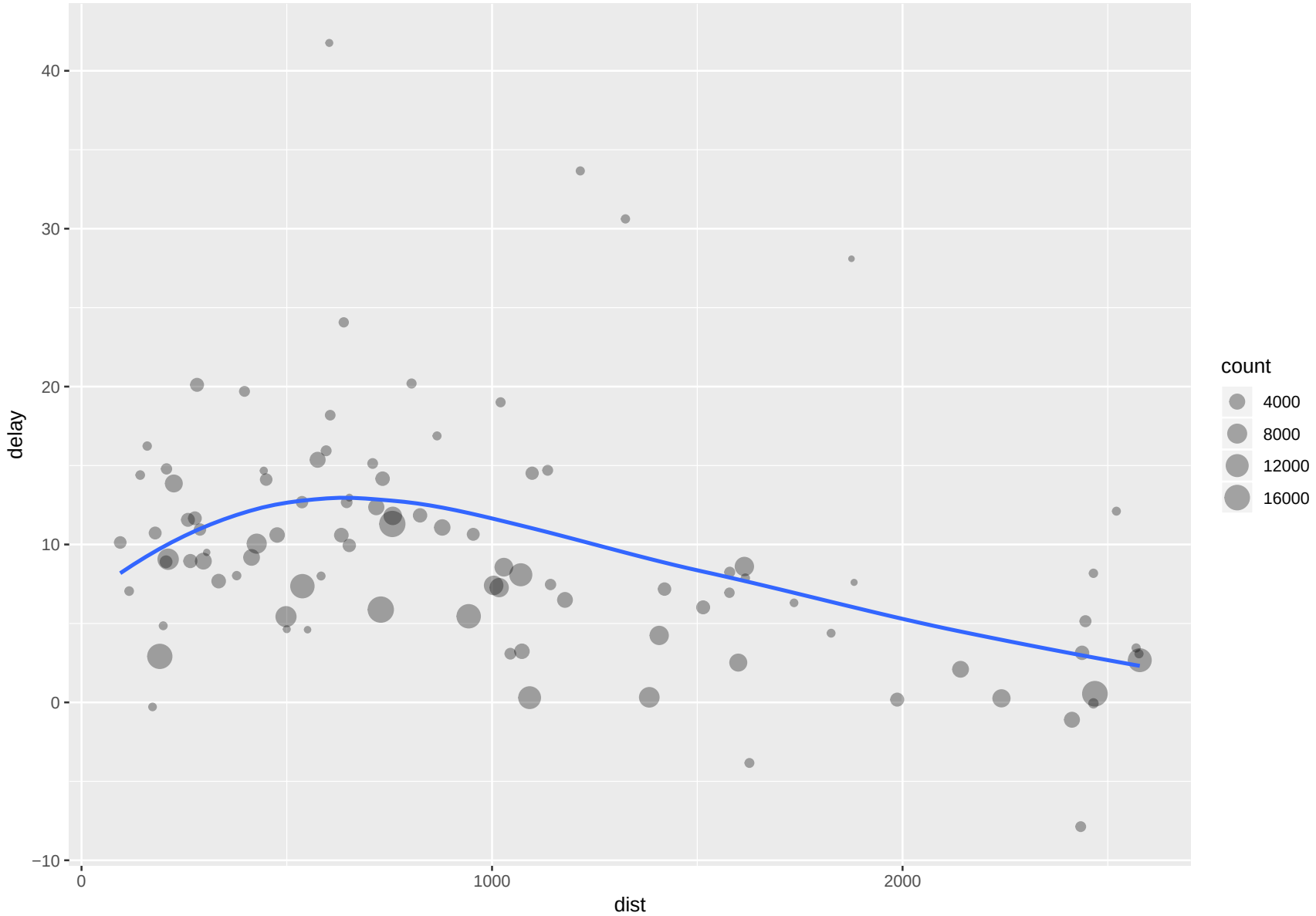
```
by_dest <- group_by(flights, dest)
delay <- summarise(by_dest,
  count = n(),
  dist = mean(distance, na.rm = TRUE),
  delay = mean(arr_delay, na.rm = TRUE)
)
delay <- filter(delay, count > 20, dest != "HNL")
```

Combining multiple operations with the pipe %>%

```
ggplot(data = delay, mapping = aes(x = dist, y = delay)) +
  geom_point(aes(size = count), alpha = 1/3) +
  geom_smooth(se = FALSE)
```

Combining multiple operations with the pipe %>%

```
## `geom_smooth()` using method = 'loess' and formula 'y ~
```



Combining multiple operations with the pipe %>%

It looks like delays increase with distance up to ~750 miles and then decrease. Maybe as flights get longer there's more ability to make up delays in the air?

> *`geom_smooth()` using method = 'loess' and formula 'y ~ x'*

Combining multiple operations with the pipe %>%

Does this code read better?

```
delays <- flights %>%
  group_by(dest) %>%
  summarise(
    count = n(),
    dist = mean(distance, na.rm = TRUE),
    delay = mean(arr_delay, na.rm = TRUE)
  ) %>%
  filter(count > 20, dest != "HNL")
```