Lab sheet 7a: Data Science Basics

Data transformation

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

dplyr basics

```
filter()
```

```
flights |>
  filter(dep_delay > 120)
## # A tibble: 9,723 x 19
##
       year month
                     day dep_time sched_dep_time dep_delay arr_time
##
      <int> <int> <int>
                                                       <dbl>
                            <int>
                                            <int>
                                                                <int>
##
   1 2013
                              848
                                                         853
                                                                 1001
                1
                       1
                                             1835
    2 2013
                                                         144
##
                1
                       1
                              957
                                              733
                                                                 1056
    3 2013
##
                1
                       1
                             1114
                                              900
                                                         134
                                                                 1447
##
   4 2013
                1
                       1
                             1540
                                             1338
                                                         122
                                                                 2020
   5 2013
                                                         290
##
                       1
                             1815
                                             1325
                                                                 2120
   6 2013
##
                       1
                             1842
                                             1422
                                                         260
                                                                 1958
                1
##
    7
       2013
                1
                       1
                             1856
                                             1645
                                                         131
                                                                 2212
##
    8
       2013
                       1
                             1934
                                             1725
                                                         129
                                                                 2126
                1
##
   9 2013
                             1938
                                             1703
                                                         155
                                                                 2109
## 10 2013
                             1942
                                             1705
                                                         157
                                                                 2124
                1
                       1
## # i 9,713 more rows
## # i 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>
# Flights that departed on January 1
flights |>
  filter(month == 1 & day == 1)
```

```
## # A tibble: 842 x 19
##
       year month
                     day dep_time sched_dep_time dep_delay arr_time
##
      <int> <int> <int>
                             <int>
                                              <int>
                                                         <dbl>
                                                                  <int>
    1 2013
                               517
                                                515
                                                             2
                                                                     830
                 1
                        1
    2
       2013
                               533
                                                529
                                                             4
                                                                     850
##
                 1
                        1
    3
       2013
                               542
                                                540
                                                             2
                                                                     923
##
                 1
                        1
    4 2013
##
                        1
                               544
                                                545
                                                            -1
                                                                   1004
##
    5 2013
                        1
                               554
                                                600
                                                            -6
                                                                    812
                 1
    6 2013
##
                 1
                        1
                               554
                                                558
                                                            -4
                                                                    740
##
    7
       2013
                        1
                               555
                                                600
                                                            -5
                                                                    913
                 1
   8 2013
                                                            -3
                                                                    709
##
                        1
                               557
                                                600
##
    9 2013
                               557
                                                600
                                                            -3
                                                                    838
```

```
558
## 10 2013
                1
                                              600
                                                         -2
                                                                  753
## # i 832 more rows
## # i 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>
# Flights that departed in November or December
flights |>
 filter(month == 11 | month == 12)
## # A tibble: 55,403 x 19
                    day dep_time sched_dep_time dep_delay arr_time
##
       year month
##
                            <int>
                                                      <dbl>
      <int> <int> <int>
                                            <int>
                                                                <int>
##
   1 2013
               11
                       1
                                5
                                             2359
                                                          6
                                                                  352
##
    2 2013
                               35
                                             2250
                                                        105
                                                                  123
               11
                       1
   3 2013
                              455
                                              500
                                                         -5
                                                                  641
##
               11
                       1
   4 2013
##
                              539
                                              545
                                                         -6
                                                                  856
               11
                       1
   5 2013
##
               11
                              542
                                              545
                                                         -3
                                                                  831
                       1
##
    6 2013
               11
                       1
                              549
                                              600
                                                        -11
                                                                  912
##
   7 2013
               11
                       1
                              550
                                              600
                                                        -10
                                                                  705
##
   8 2013
                                                                  659
               11
                       1
                              554
                                              600
                                                         -6
##
   9 2013
               11
                              554
                                              600
                                                         -6
                                                                  826
                       1
## 10 2013
               11
                       1
                              554
                                              600
                                                         -6
                                                                  749
## # i 55,393 more rows
## # i 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>
# Flights that departed in November or December: another way
flights |>
 filter(month %in% c(11,12))
## # A tibble: 55,403 x 19
       year month
                    day dep_time sched_dep_time dep_delay arr_time
##
      <int> <int> <int>
                            <int>
                                            <int>
                                                      <dbl>
                                                                <int>
##
   1 2013
               11
                       1
                                5
                                             2359
                                                          6
                                                                  352
    2 2013
                                             2250
                                                        105
                                                                  123
##
               11
                       1
                               35
##
   3 2013
               11
                       1
                              455
                                              500
                                                         -5
                                                                  641
##
   4 2013
               11
                       1
                              539
                                              545
                                                         -6
                                                                  856
   5 2013
                              542
                                                         -3
                                                                  831
##
               11
                       1
                                              545
   6 2013
##
               11
                       1
                              549
                                              600
                                                        -11
                                                                  912
   7 2013
##
               11
                       1
                              550
                                              600
                                                        -10
                                                                  705
##
   8 2013
               11
                       1
                              554
                                              600
                                                         -6
                                                                  659
    9 2013
                                                                  826
##
               11
                       1
                              554
                                              600
                                                         -6
## 10
       2013
               11
                       1
                              554
                                              600
                                                         -6
                                                                  749
## # i 55,393 more rows
## # i 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>
```

```
arrange()
```

```
flights |>
  arrange(year, month, day, dep_time)
## # A tibble: 336,776 x 19
##
       year month
                    day dep_time sched_dep_time dep_delay arr_time
##
      <int> <int> <int>
                           <int>
                                           <int>
                                                      <dbl>
##
   1 2013
                1
                      1
                              517
                                             515
                                                          2
                                                                 830
    2 2013
                              533
                                             529
                                                          4
                                                                 850
##
                1
                      1
##
   3 2013
                1
                      1
                              542
                                             540
                                                          2
                                                                 923
   4 2013
##
                      1
                             544
                                             545
                                                         -1
                                                                1004
##
  5 2013
                             554
                                             600
                                                         -6
                1
                      1
                                                                 812
## 6 2013
                1
                      1
                             554
                                             558
                                                         -4
                                                                 740
##
  7 2013
                                                         -5
                      1
                             555
                                             600
                                                                 913
                1
##
  8 2013
                1
                      1
                             557
                                             600
                                                         -3
                                                                 709
## 9 2013
                                                         -3
                                                                 838
                1
                      1
                              557
                                             600
## 10 2013
                1
                      1
                              558
                                             600
                                                         -2
                                                                 753
## # i 336,766 more rows
## # i 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>
flights |>
  arrange(desc(dep_delay))
## # A tibble: 336,776 x 19
                    day dep_time sched_dep_time dep_delay arr_time
##
       year month
##
      <int> <int> <int>
                           <int>
                                           <int>
                                                      <dbl>
   1 2013
##
                1
                      9
                             641
                                             900
                                                      1301
                                                                1242
##
    2 2013
                6
                     15
                             1432
                                            1935
                                                      1137
                                                                1607
##
  3 2013
                     10
                             1121
                                            1635
                                                      1126
                                                                1239
                1
##
  4 2013
                     20
                9
                             1139
                                            1845
                                                      1014
                                                                1457
## 5 2013
                7
                     22
                             845
                                            1600
                                                      1005
                                                                1044
##
   6 2013
                4
                     10
                             1100
                                            1900
                                                       960
                                                                1342
##
  7 2013
                3
                     17
                             2321
                                                       911
                                                                135
                                             810
##
   8 2013
                6
                     27
                             959
                                            1900
                                                       899
                                                                1236
## 9 2013
                7
                     22
                             2257
                                             759
                                                       898
                                                                 121
## 10 2013
               12
                      5
                             756
                                            1700
                                                       896
                                                                1058
## # i 336,766 more rows
## # i 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>
distinct()
# Remove duplicate rows, if any
flights |>
 distinct()
## # A tibble: 336,776 x 19
##
                    day dep_time sched_dep_time dep_delay arr_time
##
      <int> <int> <int>
                           <int>
                                           <int>
                                                      <dbl>
                                                               <int>
```

```
##
    1 2013
                       1
                              517
                                               515
                                                                   830
                1
##
    2 2013
                       1
                              533
                                               529
                                                           4
                                                                   850
                1
    3 2013
##
                       1
                              542
                                               540
                                                           2
                                                                   923
   4 2013
                                               545
##
                       1
                              544
                                                          -1
                                                                  1004
                 1
##
    5
       2013
                 1
                       1
                              554
                                               600
                                                          -6
                                                                   812
##
   6 2013
                       1
                                               558
                                                          -4
                 1
                              554
                                                                   740
##
    7 2013
                       1
                                               600
                                                          -5
                 1
                              555
                                                                   913
    8 2013
                                                          -3
##
                 1
                       1
                              557
                                               600
                                                                   709
##
    9
       2013
                 1
                       1
                               557
                                               600
                                                          -3
                                                                   838
## 10 2013
                               558
                                               600
                                                          -2
                                                                   753
                 1
                       1
## # i 336,766 more rows
## # i 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>
flights |>
 distinct(origin, dest, .keep_all = TRUE)
## # A tibble: 224 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
                              533
                                               529
                                                           4
                                                                   850
                 1
   3 2013
                                                           2
                                                                   923
##
                1
                       1
                              542
                                               540
   4 2013
                              544
                                               545
                                                          -1
##
                 1
                       1
                                                                  1004
##
   5 2013
                       1
                              554
                                               600
                                                          -6
                                                                   812
                1
##
   6 2013
                 1
                       1
                              554
                                               558
                                                          -4
                                                                   740
    7 2013
                                               600
                                                          -5
                                                                   913
##
                 1
                       1
                              555
##
    8 2013
                 1
                       1
                              557
                                               600
                                                          -3
                                                                   709
##
   9 2013
                                               600
                                                          -3
                       1
                              557
                                                                   838
                 1
## 10 2013
                 1
                       1
                              558
                                               600
                                                          -2
                                                                   753
## # i 214 more rows
## # i 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>
flights |>
  count(origin, dest, sort = TRUE)
## # A tibble: 224 x 3
##
      origin dest
##
      <chr>
             <chr> <int>
##
    1 JFK
             LAX
                    11262
##
    2 LGA
             ATL
                    10263
##
   3 LGA
                     8857
             ORD
##
    4 JFK
             SFO
                     8204
##
   5 LGA
                     6168
             CLT
    6 EWR
##
             ORD
                     6100
##
    7 JFK
             BOS
                     5898
##
    8 LGA
             MIA
                     5781
##
   9 JFK
             MCO
                     5464
## 10 EWR
             BOS
                     5327
```

i 214 more rows

```
mutate()
flights |>
  mutate(
    gain = dep_delay - arr_delay,
    speed = distance / air_time * 60
## # A tibble: 336,776 x 21
                    day dep_time sched_dep_time dep_delay arr_time
       year month
##
                                                    <dbl>
      <int> <int> <int>
                           <int>
                                          <int>
                                                             <int>
## 1 2013
                      1
                             517
                                            515
                                                        2
                                                               830
               1
## 2 2013
                      1
                             533
                                            529
                                                        4
                                                               850
## 3 2013
                             542
                                            540
                                                        2
                                                               923
                      1
                1
## 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
                                                       -5
                1
                      1
                             555
                                            600
                                                               913
## 8 2013
                1
                      1
                             557
                                            600
                                                       -3
                                                               709
## 9 2013
                                            600
                                                       -3
                                                               838
                      1
                             557
                1
## 10 2013
                             558
                                            600
                                                       -2
                                                               753
## # i 336,766 more rows
## # i 14 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>, gain <dbl>, speed <dbl>
flights |>
  mutate(
    gain = dep delay - arr delay,
    speed = distance / air_time * 60,
    .before = 1
  )
flights |>
  mutate(
    gain = dep_delay - arr_delay,
    speed = distance / air_time * 60,
    .after = day
  )
flights |>
  mutate(
    gain = dep_delay - arr_delay,
   hours = air_time / 60,
    gain_per_hour = gain / hours,
    .keep = "used"
Piping
```

```
flights |>
  filter(dest == "IAH") |>
  mutate(speed = distance / air_time * 60) |>
  select(year:day, dep_time, carrier, flight, speed) |>
```

```
## # A tibble: 7,198 x 7
##
                    day dep_time carrier flight speed
       year month
##
      <int> <int> <int>
                            <int> <chr>
                                           <int> <dbl>
##
    1 2013
                7
                      9
                              707 UA
                                             226
                                                  522.
##
    2 2013
                8
                     27
                             1850 UA
                                            1128 521.
##
   3 2013
                     28
                              902 UA
                                            1711 519.
                8
   4 2013
##
                     28
                             2122 UA
                                            1022
                                                  519.
                8
   5 2013
##
                6
                     11
                             1628 UA
                                            1178
                                                  515.
##
   6 2013
                8
                     27
                             1017 UA
                                             333
                                                  515.
##
   7 2013
                8
                     27
                             1205 UA
                                            1421
                                                  515.
   8 2013
                     27
                             1758 UA
##
                8
                                             302
                                                  515.
##
   9
       2013
                9
                     27
                              521 UA
                                             252
                                                  515.
## 10 2013
                     28
                              625 UA
                                             559 515.
                8
## # i 7,188 more rows
Groups
flights |>
  group_by(month)
## # A tibble: 336,776 x 19
               month [12]
## # Groups:
##
                    day dep_time sched_dep_time dep_delay arr_time
       year month
##
                                                      <dbl>
      <int> <int> <int>
                            <int>
                                           <int>
                                                               <int>
   1 2013
                                                          2
                                                                 830
##
                1
                      1
                              517
                                             515
##
   2 2013
                1
                      1
                              533
                                             529
                                                          4
                                                                 850
   3 2013
                                                          2
##
                1
                      1
                              542
                                             540
                                                                 923
   4 2013
                                             545
                                                                1004
##
                1
                      1
                              544
                                                         -1
##
   5 2013
                                             600
                                                         -6
                1
                      1
                              554
                                                                 812
   6 2013
##
                      1
                              554
                                             558
                                                         -4
                                                                 740
                1
##
   7 2013
                1
                      1
                              555
                                             600
                                                         -5
                                                                 913
##
   8 2013
                      1
                              557
                                             600
                                                         -3
                                                                 709
                1
##
  9 2013
                              557
                                             600
                                                         -3
                                                                 838
## 10 2013
                              558
                                             600
                                                         -2
                                                                 753
                1
                      1
## # i 336,766 more rows
## # i 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>
flights |>
  group_by(month) |>
  summarize(
    avg_delay = mean(dep_delay)
## # A tibble: 12 x 2
##
      month avg_delay
                <dbl>
##
      <int>
##
    1
          1
                   NA
   2
                   NA
##
          2
##
   3
          3
                   NA
##
          4
                   NA
    4
```

arrange(desc(speed))

```
## 5
                  NA
## 6
                  NA
         6
## 7
        7
                  NA
## 8
        8
                  NA
## 9
         9
                  NA
## 10
        10
                  NA
## 11
        11
                  NA
## 12
                  NA
        12
flights |>
 group_by(month) |>
  summarize(
   delay = mean(dep_delay, na.rm = TRUE)
## # A tibble: 12 x 2
##
     month delay
     <int> <dbl>
##
##
         1 10.0
   1
##
   2
         2 10.8
## 3
         3 13.2
## 4
         4 13.9
        5 13.0
## 5
## 6
         6 20.8
## 7
        7 21.7
## 8
        8 12.6
## 9
         9 6.72
## 10
        10 6.24
## 11
        11 5.44
## 12
        12 16.6
flights |>
 group_by(month) |>
 summarize(
   delay = mean(dep_delay, na.rm = TRUE),n=n()
 )
## # A tibble: 12 x 3
##
     month delay
##
     <int> <dbl> <int>
## 1
         1 10.0 27004
         2 10.8 24951
## 2
## 3
         3 13.2 28834
## 4
         4 13.9 28330
         5 13.0 28796
## 5
         6 20.8 28243
## 6
## 7
         7 21.7 29425
## 8
         8 12.6 29327
## 9
         9 6.72 27574
## 10
        10 6.24 28889
## 11
        11 5.44 27268
## 12
        12 16.6 28135
```