SC-MD_Sept_26-1

Srijan Kundu

2022-09-26

Working with NYC Flights Data

```
library(ggplot2)
library(dplyr)
```

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

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

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

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --

## v tibble  3.1.8     v purrr   0.3.4
## v tidyr   1.2.0     v stringr 1.4.1
## v readr   2.1.2     v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(nycflights13)
data("flights")
dim(flights)
```

```
## [1] 336776     19
```

```
head(flights)
```

```
## # A tibble: 6 x 19
##    year month   day dep_time sched_dep~1 dep_d~2 arr_t~3 sched~4 arr_d~5 carrier
##   <int> <int> <int>    <int>       <int>   <dbl>   <int>   <int>   <dbl> <chr>
## 1  2013     1     1      517         515       2     830     819      11 UA
## 2  2013     1     1      533         529       4     850     830      20 UA
## 3  2013     1     1      542         540       2     923     850      33 AA
## 4  2013     1     1      544         545      -1    1004    1022     -18 B6
## 5  2013     1     1      554         600      -6     812     837     -25 DL
## 6  2013     1     1      554         558      -4     740     728      12 UA
## # ... with 9 more variables: flight <int>, tailnum <chr>, origin <chr>,
## #   dest <chr>, air_time <dbl>, distance <dbl>, hour <dbl>, minute <dbl>,
## #   time_hour <dttm>, and abbreviated variable names 1: sched_dep_time,
## #   2: dep_delay, 3: arr_time, 4: sched_arr_time, 5: arr_delay
```

Question 1: Give us all flights departed on 1^{st} January.

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

```
## # A tibble: 842 x 19
##     year month   day dep_time sched_de~1 dep_d~2 arr_t~3 sched~4 arr_d~5 carrier
##    <int> <int> <int>    <int>      <int>   <dbl>   <int>   <int>   <dbl> <chr>
##  1  2013     1     1      517        515       2     830     819      11 UA
##  2  2013     1     1      533        529       4     850     830      20 UA
##  3  2013     1     1      542        540       2     923     850      33 AA
##  4  2013     1     1      544        545      -1    1004    1022     -18 B6
##  5  2013     1     1      554        600      -6     812     837     -25 DL
##  6  2013     1     1      554        558      -4     740     728      12 UA
##  7  2013     1     1      555        600      -5     913     854      19 B6
##  8  2013     1     1      557        600      -3     709     723     -14 EV
##  9  2013     1     1      557        600      -3     838     846      -8 B6
## 10  2013     1     1      558        600      -2     753     745       8 AA
## # ... with 832 more rows, 9 more variables: flight <int>, tailnum <chr>,
## #   origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>, hour <dbl>,
## #   minute <dbl>, time_hour <dttm>, and abbreviated variable names
## #   1: sched_dep_time, 2: dep_delay, 3: arr_time, 4: sched_arr_time,
## #   5: arr_delay
```

Question 2: Obtain all flights departed in November or December.

```
filter(flights, flights$month == 11 | flights$month == 12)
```

```
## # A tibble: 55,403 x 19
##     year month   day dep_time sched_de~1 dep_d~2 arr_t~3 sched~4 arr_d~5 carrier
##    <int> <int> <int>    <int>      <int>   <dbl>   <int>   <int>   <dbl> <chr>
##  1  2013    11     1        5       2359       6     352     345       7 B6
##  2  2013    11     1       35       2250     105     123    2356      87 B6
##  3  2013    11     1      455        500      -5     641     651     -10 US
##  4  2013    11     1      539        545      -6     856     827      29 UA
##  5  2013    11     1      542        545      -3     831     855     -24 AA
##  6  2013    11     1      549        600     -11     912     923     -11 UA
##  7  2013    11     1      550        600     -10     705     659       6 US
##  8  2013    11     1      554        600      -6     659     701      -2 US
##  9  2013    11     1      554        600      -6     826     827      -1 DL
## 10  2013    11     1      554        600      -6     749     751      -2 DL
## # ... with 55,393 more rows, 9 more variables: flight <int>, tailnum <chr>,
## #   origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>, hour <dbl>,
## #   minute <dbl>, time_hour <dttm>, and abbreviated variable names
## #   1: sched_dep_time, 2: dep_delay, 3: arr_time, 4: sched_arr_time,
## #   5: arr_delay
```

```
filter(flights, month %in% c(11, 12))
```

```
## # A tibble: 55,403 x 19
##     year month   day dep_time sched_de~1 dep_d~2 arr_t~3 sched~4 arr_d~5 carrier
##    <int> <int> <int>    <int>      <int>   <dbl>   <int>   <int>   <dbl> <chr>
##  1  2013    11     1        5       2359       6     352     345       7 B6
##  2  2013    11     1       35       2250     105     123    2356      87 B6
##  3  2013    11     1      455        500      -5     641     651     -10 US
##  4  2013    11     1      539        545      -6     856     827      29 UA
##  5  2013    11     1      542        545      -3     831     855     -24 AA
##  6  2013    11     1      549        600     -11     912     923     -11 UA
##  7  2013    11     1      550        600     -10     705     659       6 US
##  8  2013    11     1      554        600      -6     659     701      -2 US
##  9  2013    11     1      554        600      -6     826     827      -1 DL
## 10  2013    11     1      554        600      -6     749     751      -2 DL
## # ... with 55,393 more rows, 9 more variables: flight <int>, tailnum <chr>,
## #   origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>, hour <dbl>,
## #   minute <dbl>, time_hour <dttm>, and abbreviated variable names
## #   1: sched_dep_time, 2: dep_delay, 3: arr_time, 4: sched_arr_time,
## #   5: arr_delay
```

Question 3: Flights that were not delayed by more than 2 hours both for arrival or departure.

```
filter(flights, flights$dep_delay <= 120 & flights$arr_delay <= 120)
```

```
## # A tibble: 316,050 x 19
##     year month   day dep_time sched_de~1 dep_d~2 arr_t~3 sched~4 arr_d~5 carrier
##    <int> <int> <int>    <int>      <int>   <dbl>   <int>   <int>   <dbl> <chr>
##  1  2013     1     1      517        515       2     830     819      11 UA
##  2  2013     1     1      533        529       4     850     830      20 UA
##  3  2013     1     1      542        540       2     923     850      33 AA
##  4  2013     1     1      544        545      -1    1004    1022     -18 B6
##  5  2013     1     1      554        600      -6     812     837     -25 DL
##  6  2013     1     1      554        558      -4     740     728      12 UA
##  7  2013     1     1      555        600      -5     913     854      19 B6
##  8  2013     1     1      557        600      -3     709     723     -14 EV
##  9  2013     1     1      557        600      -3     838     846      -8 B6
## 10  2013     1     1      558        600      -2     753     745       8 AA
## # ... with 316,040 more rows, 9 more variables: flight <int>, tailnum <chr>,
## #   origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>, hour <dbl>,
## #   minute <dbl>, time_hour <dttm>, and abbreviated variable names
## #   1: sched_dep_time, 2: dep_delay, 3: arr_time, 4: sched_arr_time,
## #   5: arr_delay
```

```
filter(flights, !(flights$dep_delay > 120 | flights$arr_delay > 120))
```

```
## # A tibble: 316,050 x 19
##     year month   day dep_time sched_de~1 dep_d~2 arr_t~3 sched~4 arr_d~5 carrier
##    <int> <int> <int>    <int>      <int>   <dbl>   <int>   <int>   <dbl> <chr>
##  1  2013     1     1      517        515       2     830     819      11 UA
##  2  2013     1     1      533        529       4     850     830      20 UA
##  3  2013     1     1      542        540       2     923     850      33 AA
##  4  2013     1     1      544        545      -1    1004    1022     -18 B6
##  5  2013     1     1      554        600      -6     812     837     -25 DL
##  6  2013     1     1      554        558      -4     740     728      12 UA
##  7  2013     1     1      555        600      -5     913     854      19 B6
##  8  2013     1     1      557        600      -3     709     723     -14 EV
##  9  2013     1     1      557        600      -3     838     846      -8 B6
## 10  2013     1     1      558        600      -2     753     745       8 AA
## # ... with 316,040 more rows, 9 more variables: flight <int>, tailnum <chr>,
## #   origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>, hour <dbl>,
## #   minute <dbl>, time_hour <dttm>, and abbreviated variable names
## #   1: sched_dep_time, 2: dep_delay, 3: arr_time, 4: sched_arr_time,
## #   5: arr_delay
```

Question 4: Flights with arrival delay of 2 or more hours

```
filter(flights, flights$arr_delay >= 120)
```

```
## # A tibble: 10,200 x 19
##     year month   day dep_time sched_de~1 dep_d~2 arr_t~3 sched~4 arr_d~5 carrier
##    <int> <int> <int>    <int>      <int>   <dbl>   <int>   <int>   <dbl> <chr>
##  1  2013     1     1      811        630     101    1047     830     137 MQ
##  2  2013     1     1      848       1835     853    1001    1950     851 MQ
##  3  2013     1     1      957        733     144    1056     853     123 UA
##  4  2013     1     1     1114        900     134    1447    1222     145 UA
##  5  2013     1     1     1505       1310     115    1638    1431     127 EV
##  6  2013     1     1     1525       1340     105    1831    1626     125 B6
##  7  2013     1     1     1549       1445      64    1912    1656     136 EV
##  8  2013     1     1     1558       1359     119    1718    1515     123 EV
##  9  2013     1     1     1732       1630      62    2028    1825     123 EV
## 10  2013     1     1     1803       1620     103    2008    1750     138 MQ
## # ... with 10,190 more rows, 9 more variables: flight <int>, tailnum <chr>,
## #   origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>, hour <dbl>,
## #   minute <dbl>, time_hour <dttm>, and abbreviated variable names
## #   1: sched_dep_time, 2: dep_delay, 3: arr_time, 4: sched_arr_time,
## #   5: arr_delay
```

Question 5: Flights that flew to Houston

```
filter(flights, dest %in% c('IAH', 'HOU', 'EFD'))
```

```
## # A tibble: 9,313 x 19
##     year month   day dep_time sched_de~1 dep_d~2 arr_t~3 sched~4 arr_d~5 carrier
##    <int> <int> <int>    <int>      <int>   <dbl>   <int>   <int>   <dbl> <chr>
##  1  2013     1     1      517        515       2     830     819      11 UA
##  2  2013     1     1      533        529       4     850     830      20 UA
##  3  2013     1     1      623        627      -4     933     932       1 UA
##  4  2013     1     1      728        732      -4    1041    1038       3 UA
##  5  2013     1     1      739        739       0    1104    1038      26 UA
##  6  2013     1     1      908        908       0    1228    1219       9 UA
##  7  2013     1     1     1028       1026       2    1350    1339      11 UA
##  8  2013     1     1     1044       1045      -1    1352    1351       1 UA
##  9  2013     1     1     1114        900     134    1447    1222     145 UA
## 10  2013     1     1     1205       1200       5    1503    1505      -2 UA
## # ... with 9,303 more rows, 9 more variables: flight <int>, tailnum <chr>,
## #   origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>, hour <dbl>,
## #   minute <dbl>, time_hour <dttm>, and abbreviated variable names
## #   1: sched_dep_time, 2: dep_delay, 3: arr_time, 4: sched_arr_time,
## #   5: arr_delay
```

Question 6: Flights operated by United, American or Delta.

```
filter(flights, carrier %in% c('UA', 'AA', 'DL'))
```

```
## # A tibble: 139,504 x 19
##     year month   day dep_time sched_de~1 dep_d~2 arr_t~3 sched~4 arr_d~5 carrier
##    <int> <int> <int>    <int>      <int>   <dbl>   <int>   <int>   <dbl> <chr>
##  1  2013     1     1      517        515       2     830     819      11 UA
##  2  2013     1     1      533        529       4     850     830      20 UA
##  3  2013     1     1      542        540       2     923     850      33 AA
##  4  2013     1     1      554        600      -6     812     837     -25 DL
##  5  2013     1     1      554        558      -4     740     728      12 UA
##  6  2013     1     1      558        600      -2     753     745       8 AA
##  7  2013     1     1      558        600      -2     924     917       7 UA
##  8  2013     1     1      558        600      -2     923     937     -14 UA
##  9  2013     1     1      559        600      -1     941     910      31 AA
## 10  2013     1     1      559        600      -1     854     902      -8 UA
## # ... with 139,494 more rows, 9 more variables: flight <int>, tailnum <chr>,
## #   origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>, hour <dbl>,
## #   minute <dbl>, time_hour <dttm>, and abbreviated variable names
## #   1: sched_dep_time, 2: dep_delay, 3: arr_time, 4: sched_arr_time,
## #   5: arr_delay
```

Question 7: Flights that departed in summer (July, August, September)

```
filter(flights, month %in% c(7, 8, 9))
```

```
## # A tibble: 86,326 x 19
##     year month   day dep_time sched_de~1 dep_d~2 arr_t~3 sched~4 arr_d~5 carrier
##    <int> <int> <int>    <int>      <int>   <dbl>   <int>   <int>   <dbl> <chr>
##  1  2013     7     1        1       2029     212     236    2359     157 B6
##  2  2013     7     1        2       2359       3     344     344       0 B6
##  3  2013     7     1       29       2245     104     151       1     110 B6
##  4  2013     7     1       43       2130     193     322      14     188 B6
##  5  2013     7     1       44       2150     174     300     100     120 AA
##  6  2013     7     1       46       2051     235     304    2358     186 B6
##  7  2013     7     1       48       2001     287     308    2305     243 VX
##  8  2013     7     1       58       2155     183     335      43     172 B6
##  9  2013     7     1      100       2146     194     327      30     177 B6
## 10  2013     7     1      100       2245     135     337     135     122 B6
## # ... with 86,316 more rows, 9 more variables: flight <int>, tailnum <chr>,
## #   origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>, hour <dbl>,
## #   minute <dbl>, time_hour <dttm>, and abbreviated variable names
## #   1: sched_dep_time, 2: dep_delay, 3: arr_time, 4: sched_arr_time,
## #   5: arr_delay
```

Question 8: Flights that arrived more than 2 hours late but did not leave late.

```
filter(flights, flights$arr_delay > 120 & flights$dep_delay <= 0)
```

```
## # A tibble: 29 x 19
##     year month   day dep_time sched_de~1 dep_d~2 arr_t~3 sched~4 arr_d~5 carrier
##    <int> <int> <int>    <int>      <int>   <dbl>   <int>   <int>   <dbl> <chr>
##  1  2013     1    27     1419       1420      -1    1754    1550     124 MQ
##  2  2013    10     7     1350       1350       0    1736    1526     130 EV
##  3  2013    10     7     1357       1359      -2    1858    1654     124 AA
##  4  2013    10    16      657        700      -3    1258    1056     122 B6
##  5  2013    11     1      658        700      -2    1329    1015     194 VX
##  6  2013     3    18     1844       1847      -3      39    2219     140 UA
##  7  2013     4    17     1635       1640      -5    2049    1845     124 MQ
##  8  2013     4    18      558        600      -2    1149     850     179 AA
##  9  2013     4    18      655        700      -5    1213     950     143 AA
## 10  2013     5    22     1827       1830      -3    2217    2010     127 MQ
## # ... with 19 more rows, 9 more variables: flight <int>, tailnum <chr>,
## #   origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>, hour <dbl>,
## #   minute <dbl>, time_hour <dttm>, and abbreviated variable names
## #   1: sched_dep_time, 2: dep_delay, 3: arr_time, 4: sched_arr_time,
## #   5: arr_delay
```

Question 9: Flights that were delayed by at least an hour, but made up over 30 mins in flight.

```
filter(flights, dep_delay >= 60,(dep_delay-arr_delay > 30))
```

```
## # A tibble: 1,844 x 19
##     year month   day dep_time sched_de~1 dep_d~2 arr_t~3 sched~4 arr_d~5 carrier
##    <int> <int> <int>    <int>      <int>   <dbl>   <int>   <int>   <dbl> <chr>
##  1  2013     1     1     2205       1720     285      46    2040     246 AA
##  2  2013     1     1     2326       2130     116     131      18      73 B6
##  3  2013     1     3     1503       1221     162    1803    1555     128 UA
##  4  2013     1     3     1839       1700      99    2056    1950      66 AA
##  5  2013     1     3     1850       1745      65    2148    2120      28 AA
##  6  2013     1     3     1941       1759     102    2246    2139      67 UA
##  7  2013     1     3     1950       1845      65    2228    2227       1 B6
##  8  2013     1     3     2015       1915      60    2135    2111      24 9E
##  9  2013     1     3     2257       2000     177      45    2224     141 9E
## 10  2013     1     4     1917       1700     137    2135    1950     105 AA
## # ... with 1,834 more rows, 9 more variables: flight <int>, tailnum <chr>,
## #   origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>, hour <dbl>,
## #   minute <dbl>, time_hour <dttm>, and abbreviated variable names
## #   1: sched_dep_time, 2: dep_delay, 3: arr_time, 4: sched_arr_time,
## #   5: arr_delay
```

Question 10: Flights that departed between midnight and 6:00 am, both inclusive.

```
filter(flights, dep_time >= 0000 & dep_time <= 0600)
```

```
## # A tibble: 9,344 x 19
##     year month   day dep_time sched_de~1 dep_d~2 arr_t~3 sched~4 arr_d~5 carrier
##    <int> <int> <int>    <int>      <int>   <dbl>   <int>   <int>   <dbl> <chr>
##  1  2013     1     1      517        515       2     830     819      11 UA
##  2  2013     1     1      533        529       4     850     830      20 UA
##  3  2013     1     1      542        540       2     923     850      33 AA
##  4  2013     1     1      544        545      -1    1004    1022     -18 B6
##  5  2013     1     1      554        600      -6     812     837     -25 DL
##  6  2013     1     1      554        558      -4     740     728      12 UA
##  7  2013     1     1      555        600      -5     913     854      19 B6
##  8  2013     1     1      557        600      -3     709     723     -14 EV
##  9  2013     1     1      557        600      -3     838     846      -8 B6
## 10  2013     1     1      558        600      -2     753     745       8 AA
## # ... with 9,334 more rows, 9 more variables: flight <int>, tailnum <chr>,
## #   origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>, hour <dbl>,
## #   minute <dbl>, time_hour <dttm>, and abbreviated variable names
## #   1: sched_dep_time, 2: dep_delay, 3: arr_time, 4: sched_arr_time,
## #   5: arr_delay
```

```
filter(flights, between(dep_time, 0000, 600))
```

```
## # A tibble: 9,344 x 19
##     year month   day dep_time sched_de~1 dep_d~2 arr_t~3 sched~4 arr_d~5 carrier
##    <int> <int> <int>    <int>      <int>   <dbl>   <int>   <int>   <dbl> <chr>
##  1  2013     1     1      517        515       2     830     819      11 UA
##  2  2013     1     1      533        529       4     850     830      20 UA
##  3  2013     1     1      542        540       2     923     850      33 AA
##  4  2013     1     1      544        545      -1    1004    1022     -18 B6
##  5  2013     1     1      554        600      -6     812     837     -25 DL
##  6  2013     1     1      554        558      -4     740     728      12 UA
##  7  2013     1     1      555        600      -5     913     854      19 B6
##  8  2013     1     1      557        600      -3     709     723     -14 EV
##  9  2013     1     1      557        600      -3     838     846      -8 B6
## 10  2013     1     1      558        600      -2     753     745       8 AA
## # ... with 9,334 more rows, 9 more variables: flight <int>, tailnum <chr>,
## #   origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>, hour <dbl>,
## #   minute <dbl>, time_hour <dttm>, and abbreviated variable names
## #   1: sched_dep_time, 2: dep_delay, 3: arr_time, 4: sched_arr_time,
## #   5: arr_delay
```

```
filter(flights, !between(dep_time, 0601, 2359))
```

```
## # A tibble: 9,373 x 19
##     year month   day dep_time sched_de~1 dep_d~2 arr_t~3 sched~4 arr_d~5 carrier
##    <int> <int> <int>    <int>      <int>   <dbl>   <int>   <int>   <dbl> <chr>
##  1  2013     1     1      517        515       2     830     819      11 UA
##  2  2013     1     1      533        529       4     850     830      20 UA
##  3  2013     1     1      542        540       2     923     850      33 AA
##  4  2013     1     1      544        545      -1    1004    1022     -18 B6
##  5  2013     1     1      554        600      -6     812     837     -25 DL
##  6  2013     1     1      554            558        -4      740      728        12 UA
##  7  2013     1     1      555            600        -5      913      854        19 B6
##  8  2013     1     1      557            600        -3      709      723       -14 EV
##  9  2013     1     1      557            600        -3      838      846        -8 B6
## 10  2013     1     1      558            600        -2      753      745         8 AA
## # ... with 9,363 more rows, 9 more variables: flight <int>, tailnum <chr>,
## #   origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>, hour <dbl>,
## #   minute <dbl>, time_hour <dttm>, and abbreviated variable names
## #   1: sched_dep_time, 2: dep_delay, 3: arr_time, 4: sched_arr_time,
## #   5: arr_delay
```

Question 11: How many flights have a missing dep_time?

```
count(filter(flights, is.na(dep_time)))
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1  8255
```

Question 12: From the flight dataset, only consider the data on arrival delay, departure delay, distance and air time.

```
select(flights, arr_delay, dep_delay, distance, air_time)
```

```
## # A tibble: 336,776 x 4
##    arr_delay dep_delay distance air_time
##        <dbl>     <dbl>    <dbl>    <dbl>
##  1        11         2     1400      227
##  2        20         4     1416      227
##  3        33         2     1089      160
##  4       -18        -1     1576      183
##  5       -25        -6      762      116
##  6        12        -4      719      150
##  7        19        -5     1065      158
##  8       -14        -3      229       53
##  9        -8        -3      944      140
## 10         8        -2      733      138
## # ... with 336,766 more rows
```

```
flights = flights %>% mutate(gain = arr_delay - dep_delay)
flights = flights %>% mutate(speed = distance/arr_time*60)
```

Question 13: How can you compute hours and minutes from departure time using `transmute`?

```
transmute(flights,
  dep_time,
  hour = dep_time %/% 100,
  minute = dep_time %% 100
)
```

```
## # A tibble: 336,776 x 3
##    dep_time  hour minute
##       <int> <dbl>  <dbl>
##  1      517     5     17
##  2      533     5     33
##  3      542     5     42
##  4      544     5     44
##  5      554     5     54
##  6      554     5     54
##  7      555     5     55
##  8      557     5     57
##  9      557     5     57
## 10      558     5     58
## # ... with 336,766 more rows
```

Click for reference