Chapter 3 - Data Transformation with dplyr

Thalles Quinaglia Liduares

18/02/2022

Exercise 1

Upload packages

```
library(dplyr)
library(nycflights13)
```

Upload database

```
data<-nycflights13::flights

View(data)
```

1. Find all flights that:

a. Had an arrival delay of two or more hours

```
del<-data %>%
    filter(arr_delay >=120)

del
```

```
## # A tibble: 10,200 x 19
##     year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##    <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
## 1   2013     1     1      811            630       101     1047            830
## 2   2013     1     1      848           1835       853     1001           1950
## 3   2013     1     1      957            733       144     1056            853
## 4   2013     1     1     1114            900       134     1447           1222
## 5   2013     1     1     1505           1310       115     1638           1431
## 6   2013     1     1     1525           1340       105     1831           1626
## 7   2013     1     1     1549           1445        64     1912           1656
## 8   2013     1     1     1558           1359       119     1718           1515
## 9   2013     1     1     1732           1630        62     2028           1825
## 10  2013     1     1     1803           1620       103     2008           1750
## # ... with 10,190 more rows, and 11 more variables: 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>
```

b. Flew to Houston (IAH or HOU)

```
flight<-data %>%
        filter(dest =="IAH" | dest == "HOU")

flight
```

```
## # A tibble: 9,313 x 19
##     year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##    <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
## 1   2013     1     1      517            515         2      830            819
## 2   2013     1     1      533            529         4      850            830
## 3   2013     1     1      623            627        -4      933            932
## 4   2013     1     1      728            732        -4     1041           1038
## 5   2013     1     1      739            739         0     1104           1038
## 6   2013     1     1      908            908         0     1228           1219
## 7   2013     1     1     1028           1026         2     1350           1339
## 8   2013     1     1     1044           1045        -1     1352           1351
## 9   2013     1     1     1114            900       134     1447           1222
## 10  2013     1     1     1205           1200         5     1503           1505
## # ... with 9,303 more rows, and 11 more variables: 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>
```

c. Were operated by United, American, or Delta

```
op<-data %>%
    filter(carrier %in% c("UA","AA", "DL"))

op
```

```
## # A tibble: 139,504 x 19
##     year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##    <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
## 1   2013     1     1      517            515         2      830            819
## 2   2013     1     1      533            529         4      850            830
## 3   2013     1     1      542            540         2      923            850
## 4   2013     1     1      554            600        -6      812            837
## 5   2013     1     1      554            558        -4      740            728
## 6   2013     1     1      558            600        -2      753            745
## 7   2013     1     1      558            600        -2      924            917
## 8   2013     1     1      558            600        -2      923            937
## 9   2013     1     1      559            600        -1      941            910
## 10  2013     1     1      559            600        -1      854            902
## # ... with 139,494 more rows, and 11 more variables: 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>
```

d. Departed in summer (July, August, and September)

```
dpt<-data %>%
    filter(month %in% c(7,8,9))

dpt
```

```
## # A tibble: 86,326 x 19
##     year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##    <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
##  1  2013     7     1        1           2029       212      236           2359
##  2  2013     7     1        2           2359         3      344            344
##  3  2013     7     1       29           2245       104      151              1
##  4  2013     7     1       43           2130       193      322             14
##  5  2013     7     1       44           2150       174      300            100
##  6  2013     7     1       46           2051       235      304           2358
##  7  2013     7     1       48           2001       287      308           2305
##  8  2013     7     1       58           2155       183      335             43
##  9  2013     7     1      100           2146       194      327             30
## 10  2013     7     1      100           2245       135      337            135
## # ... with 86,316 more rows, and 11 more variables: 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>
```

e. Arrived more than two hours late, but didn't leave late

```
arr2<-data %>%
     filter(arr_delay > 120 & dep_delay < 0)

arr2
```

```
## # A tibble: 26 x 19
##     year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##    <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
##  1  2013     1    27     1419           1420        -1     1754           1550
##  2  2013    10     7     1357           1359        -2     1858           1654
##  3  2013    10    16      657            700        -3     1258           1056
##  4  2013    11     1      658            700        -2     1329           1015
##  5  2013     3    18     1844           1847        -3       39           2219
##  6  2013     4    17     1635           1640        -5     2049           1845
##  7  2013     4    18      558            600        -2     1149            850
##  8  2013     4    18      655            700        -5     1213            950
##  9  2013     5    22     1827           1830        -3     2217           2010
## 10  2013     6     5     1604           1615       -11     2041           1840
## # ... with 16 more rows, and 11 more variables: 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>
```

f. Were delayed by at least an hour, but made up over 30 minutes in flight

```
del2<-data %>%
     filter(arr_delay >= 60 & arr_delay <=90)

del2
```

```
## # A tibble: 11,793 x 19
##     year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##    <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
## 1   2013     1     1     1120            944        96     1331           1213
## 2   2013     1     1     1255           1200        55     1451           1330
## 3   2013     1     1     1337           1220        77     1649           1531
## 4   2013     1     1     1342           1320        22     1617           1504
## 5   2013     1     1     1356           1259        57     1538           1438
## 6   2013     1     1     1402           1323        39     1650           1526
## 7   2013     1     1     1411           1315        56     1717           1611
## 8   2013     1     1     1424           1349        35     1701           1556
## 9   2013     1     1     1428           1329        59     1803           1640
## 10  2013     1     1     1520           1430        50     1835           1735
## # ... with 11,783 more rows, and 11 more variables: 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>
```