R_dplyr

Ulises Jose Bustamante Mora

2023-10-31

Libraries

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

Dataset

```
df = flights
head(flights)
```

```
## # A tibble: 6 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      544            545        -1     1004
1022
## 5  2013     1     1      554            600        -6      812
837
## 6  2013     1     1      554            558        -4      740
728
## # i 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>
```

dplyr

Filter

```
df %>% filter(month == 5, day == 4, carrier == "AA")

## # A tibble: 76 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     5     4      541            540         1      828
840
##  2  2013     5     4      548            600       -12      831
850
##  3  2013     5     4      600            605        -5      854
910
##  4  2013     5     4      611            615        -4      904
915
##  5  2013     5     4      623            630        -7      745
805
##  6  2013     5     4      640            640         0     1023
1040
##  7  2013     5     4      652            655        -3      939
935
##  8  2013     5     4      653            700        -7      958
1010
##  9  2013     5     4      657            700        -3      918
945
## 10  2013     5     4      717            725        -8      822
905
## # i 66 more rows
## # i 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>
```

Slice

```
df %>% slice(1:7)

## # A tibble: 7 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      544            545        -1     1004
1022
## 5  2013     1     1      554            600        -6      812
837
## 6  2013     1     1      554            558        -4      740
728
## 7  2013     1     1      555            600        -5      913
854
## # i 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>
```

Arrange

```
df %>% arrange(year, desc(month), day, arr_time) %>% head()
```

```
## # A tibble: 6 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    12     1     2255           2250         5        1
2356
## 2  2013    12     1     2242           2250        -8        4
8
## 3  2013    12     1     2134           2140        -6        8
36
## 4  2013    12     1     2209           2125        44        8
2357
## 5  2013    12     1     2027           2019         8       10
2355
## 6  2013    12     1     2111           2040        31       11
2329
## # i 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>
```

Select

```
df %>% select(carrier, month) %>% head()
```

```
## # A tibble: 6 x 2
##   carrier month
##   <chr>   <int>
## 1 UA          1
## 2 UA          1
## 3 AA          1
## 4 B6          1
## 5 DL          1
## 6 UA          1
```

```
df %>% select(starts_with("dep")) %>% head(2)
```

```
## # A tibble: 2 x 2
##   dep_time dep_delay
##      <int>     <dbl>
## 1      517         2
## 2      533         4
```

```
df %>% select(ends_with("time")) %>% head(2)
```

```
## # A tibble: 2 x 5
##   dep_time sched_dep_time arr_time sched_arr_time air_time
##      <int>          <int>    <int>          <int>    <dbl>
## 1      517            515      830            819      227
## 2      533            529      850            830      227
```

Rename

```
df %>% rename(new_name = carrier) %>% select(new_name) %>% head()
```

```
## # A tibble: 6 x 1
##   new_name
##   <chr>
## 1 UA
## 2 UA
## 3 AA
## 4 B6
## 5 DL
## 6 UA
```

Distinct

```
df %>% distinct(month)
```

```
## # A tibble: 12 x 1
##    month
##    <int>
##  1     1
##  2    10
##  3    11
##  4    12
##  5     2
##  6     3
##  7     4
##  8     5
##  9     6
## 10     7
## 11     8
## 12     9
```

Mutate

```
df %>% mutate(new_col = arr_delay-dep_delay) %>% select(new_col) %>%
head()
```

```
## # A tibble: 6 x 1
##   new_col
##     <dbl>
## 1       9
## 2      16
## 3      31
## 4     -17
## 5     -19
## 6      16
```

Trasmute

```
df %>% transmute(new_col = arr_delay-dep_delay) %>% head()

## # A tibble: 6 x 1
##   new_col
##     <dbl>
## 1       9
## 2      16
## 3      31
## 4     -17
## 5     -19
## 6      16
```

Summarise

```
df %>% summarise(avg_air_time = mean(air_time, na.rm = T)) %>%
  select(avg_air_time) %>% head()

## # A tibble: 1 x 1
##   avg_air_time
##          <dbl>
## 1         151.
```

Group by

```
df %>%
  group_by(origin)

## # A tibble: 336,776 x 19
## # Groups:   origin [3]
##     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      544            545        -1     1004
1022
```

```
##   5  2013     1     1      554            600        -6      812
837
##   6  2013     1     1      554            558        -4      740
728
##   7  2013     1     1      555            600        -5      913
854
##   8  2013     1     1      557            600        -3      709
723
##   9  2013     1     1      557            600        -3      838
846
## 10  2013     1     1      558            600        -2      753
745
## # i 336,766 more rows
## # i 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>
```

Sample N

```
df %>% sample_n(5)
```

```
## # A tibble: 5 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     6    11     1010           1005         5     1238
1249
## 2  2013     6    26     1618           1559        19     1920
1914
## 3  2013     2    17     1514           1516        -2     1756
1812
## 4  2013     4     8      951            842        69     1143
1054
## 5  2013    10    14      934            931         3     1242
1235
## # i 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>
```

Sample Frac

```
df %>% sample_frac(0.2)
```

```
## # A tibble: 67,355 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     8    10     1800           1805        -5     1904
```

```
1932
##  2  2013    11    21     2028           2028         0     2149
2206
##  3  2013     6    28      540            540         0      759
807
##  4  2013     6    13     1546           1545         1     1808
1806
##  5  2013     9    18     1925           1930        -5     2154
2234
##  6  2013    12    26     1011            929        42     1115
1045
##  7  2013     4     7      841            850        -9     1129
1158
##  8  2013     8    30     1446           1454        -8     1643
1710
##  9  2013     1    17     1954           2000        -6     2305
2305
## 10  2013     2    23      828            829        -1      944
939
## # i 67,345 more rows
## # i 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>
```