Data Manipulation With Dplyr

Adejumo Ridwan Suleiman

2022-09-25

Installing dplyr

```
#install.packages("dplyr")
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
```

Installing nycflights13

This data is an RDMS: a Relational Database Management System, it is made up of more than one table of data which are related to each other. - flights - airlines - airport - planes - weather

```
#install.packages("nycflights13)
library(nycflights13)
```

```
## Warning: package 'nycflights13' was built under R version 4.2.1
```

Flights (Main Data)

Details of all flights in the year 2013

```
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
```

Airlines Names

```
head(airlines)
```

```
## # A tibble: 6 x 2
##   carrier name
##   <chr>   <chr>
## 1 9E      Endeavor Air Inc.
## 2 AA      American Airlines Inc.
## 3 AS      Alaska Airlines Inc.
## 4 B6      JetBlue Airways
## 5 DL      Delta Air Lines Inc.
## 6 EV      ExpressJet Airlines Inc.
```

Airport Metadata

```
head(airports)
```

```
## # A tibble: 6 x 8
##   faa   name                            lat   lon   alt    tz dst   tzone
##   <chr> <chr>                         <dbl> <dbl> <dbl> <dbl> <chr> <chr>
## 1 04G   Lansdowne Airport              41.1 -80.6  1044    -5 A     America/Ne~
## 2 06A   Moton Field Municipal Airport  32.5 -85.7   264    -6 A     America/Ch~
## 3 06C   Schaumburg Regional            42.0 -88.1   801    -6 A     America/Ch~
## 4 06N   Randall Airport                41.4 -74.4   523    -5 A     America/Ne~
## 5 09J   Jekyll Island Airport          31.1 -81.4    11    -5 A     America/Ne~
## 6 0A9   Elizabethton Municipal Airport 36.4 -82.2  1593    -5 A     America/Ne~
```

Planes Metadata

```
head(planes)
```

```
## # A tibble: 6 x 9
##   tailnum  year type                  manuf~1 model engines seats speed engine
##   <chr>   <int> <chr>                 <chr>   <chr>   <int> <int> <int> <chr>
## 1 N10156   2004 Fixed wing multi engine EMBRAER EMB-~       2    55    NA Turbo~
```

```
## 2 N102UW   1998 Fixed wing multi engine AIRBUS~ A320~       2   182    NA Turbo~
## 3 N103US   1999 Fixed wing multi engine AIRBUS~ A320~       2   182    NA Turbo~
## 4 N104UW   1999 Fixed wing multi engine AIRBUS~ A320~       2   182    NA Turbo~
## 5 N10575   2002 Fixed wing multi engine EMBRAER EMB-~       2    55    NA Turbo~
## 6 N105UW   1999 Fixed wing multi engine AIRBUS~ A320~       2   182    NA Turbo~
## # ... with abbreviated variable name 1: manufacturer
```

Weather (hourly)

```
head(weather)
```

```
## # A tibble: 6 x 15
##   origin  year month   day  hour  temp  dewp humid wind_dir wind_speed wind_gust
##   <chr>  <int> <int> <int> <int> <dbl> <dbl> <dbl>    <dbl>      <dbl>     <dbl>
## 1 EWR     2013     1     1     1  39.0  26.1  59.4      270      10.4         NA
## 2 EWR     2013     1     1     2  39.0  27.0  61.6      250       8.06        NA
## 3 EWR     2013     1     1     3  39.0  28.0  64.4      240      11.5         NA
## 4 EWR     2013     1     1     4  39.9  28.0  62.2      250      12.7         NA
## 5 EWR     2013     1     1     5  39.0  28.0  64.4      260      12.7         NA
## 6 EWR     2013     1     1     6  37.9  28.0  67.2      240      11.5         NA
## # ... with 4 more variables: precip <dbl>, pressure <dbl>, visib <dbl>,
## #   time_hour <dttm>
```

```
?weather
```

```
## starting httpd help server ... done
```

Grouping and Summarizing

```
month_delay <- flights |>
  group_by(month) |>
  summarize(avg_dep_delay = mean(dep_delay, na.rm = TRUE),
            avg_arr_delay = mean(arr_delay, na.rm = TRUE))
```

```
carrier_delay <- flights |>
  group_by(carrier) |>
  summarize(avg_carr_dep_delay = mean(dep_delay, na.rm = TRUE),
            avg_carr_arr_delay = mean(arr_delay, na.rm = TRUE))
```

Arranging

```
carrier_delay |>
  arrange(desc(avg_carr_dep_delay))
```

```
## # A tibble: 16 x 3
##    carrier avg_carr_dep_delay avg_carr_arr_delay
##    <chr>                <dbl>              <dbl>
##  1 F9                   20.2              21.9
##  2 EV                   20.0              15.8
##  3 YV                   19.0              15.6
##  4 FL                   18.7              20.1
##  5 WN                   17.7               9.65
##  6 9E                   16.7               7.38
##  7 B6                   13.0               9.46
##  8 VX                   12.9               1.76
##  9 OO                   12.6              11.9
## 10 UA                   12.1               3.56
## 11 MQ                   10.6              10.8
## 12 DL                    9.26              1.64
## 13 AA                    8.59              0.364
## 14 AS                    5.80             -9.93
## 15 HA                    4.90             -6.92
## 16 US                    3.78              2.13
```

Filtering

```
flights |>
  filter(month == 1 & dep_delay < 0 & arr_delay < 0)
```

```
## # A tibble: 11,491 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      544        545      -1    1004    1022     -18 B6
##  2  2013     1     1      554        600      -6     812     837     -25 DL
##  3  2013     1     1      557        600      -3     709     723     -14 EV
##  4  2013     1     1      557        600      -3     838     846      -8 B6
##  5  2013     1     1      558        600      -2     849     851      -2 B6
##  6  2013     1     1      558        600      -2     853     856      -3 B6
##  7  2013     1     1      558        600      -2     923     937     -14 UA
##  8  2013     1     1      559        600      -1     854     902      -8 UA
##  9  2013     1     1      602        610      -8     812     820      -8 DL
## 10  2013     1     1      606        610      -4     858     910     -12 AA
## # ... with 11,481 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
```

```
carrier_delay
```

```
## # A tibble: 16 x 3
##    carrier avg_carr_dep_delay avg_carr_arr_delay
##    <chr>                <dbl>              <dbl>
##  1 9E                   16.7               7.38
```

```
##  2 AA                    8.59             0.364
##  3 AS                    5.80            -9.93
##  4 B6                   13.0              9.46
##  5 DL                    9.26             1.64
##  6 EV                   20.0             15.8
##  7 F9                   20.2             21.9
##  8 FL                   18.7             20.1
##  9 HA                    4.90            -6.92
## 10 MQ                   10.6             10.8
## 11 OO                   12.6             11.9
## 12 UA                   12.1              3.56
## 13 US                    3.78             2.13
## 14 VX                   12.9              1.76
## 15 WN                   17.7              9.65
## 16 YV                   19.0             15.6
```

Selecting

```
flights |>
  select(!(hour:time_hour))
```

```
## # A tibble: 336,776 x 16
##     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 336,766 more rows, 6 more variables: flight <int>, tailnum <chr>,
## #   origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>, and abbreviated
## #   variable names 1: sched_dep_time, 2: dep_delay, 3: arr_time,
## #   4: sched_arr_time, 5: arr_delay
```

Creating Variables

```
carrier_speed <- flights |>
  mutate(speed = distance/(air_time/60)) |>
  select(carrier, speed) |>
  group_by(carrier) |>
  summarize(avg_speed = mean(speed, na.rm = TRUE)) |>
  arrange(desc(avg_speed))
```

```
flights |>
  mutate(speed = distance/(air_time/60), .keep = "all")
```

```
## # A tibble: 336,776 x 20
##     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 336,766 more rows, 10 more variables: flight <int>, tailnum <chr>,
## #   origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>, hour <dbl>,
## #   minute <dbl>, time_hour <dttm>, speed <dbl>, and abbreviated variable names
## #   1: sched_dep_time, 2: dep_delay, 3: arr_time, 4: sched_arr_time,
## #   5: arr_delay
```

Renaming

```
flights |>
  rename(destination = dest)
```

```
## # A tibble: 336,776 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 336,766 more rows, 9 more variables: flight <int>, tailnum <chr>,
## #   origin <chr>, destination <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
```

Mutating Joins

```
top5_carrier_speed <- carrier_speed |>
  head(5)
top5_carrier_speed
```

```
## # A tibble: 5 x 2
##   carrier avg_speed
##   <chr>       <dbl>
## 1 HA           480.
## 2 VX           446.
## 3 AS           444.
## 4 F9           425.
## 5 UA           421.
```

```
full_join(x = top5_carrier_speed,
          y = airlines,
          by = "carrier")
```

```
## # A tibble: 16 x 3
##    carrier avg_speed name
##    <chr>       <dbl> <chr>
##  1 HA           480. Hawaiian Airlines Inc.
##  2 VX           446. Virgin America
##  3 AS           444. Alaska Airlines Inc.
##  4 F9           425. Frontier Airlines Inc.
##  5 UA           421. United Air Lines Inc.
##  6 9E            NA  Endeavor Air Inc.
##  7 AA            NA  American Airlines Inc.
##  8 B6            NA  JetBlue Airways
##  9 DL            NA  Delta Air Lines Inc.
## 10 EV            NA  ExpressJet Airlines Inc.
## 11 FL            NA  AirTran Airways Corporation
## 12 MQ            NA  Envoy Air
## 13 OO            NA  SkyWest Airlines Inc.
## 14 US            NA  US Airways Inc.
## 15 WN            NA  Southwest Airlines Co.
## 16 YV            NA  Mesa Airlines Inc.
```

```
inner_join(x = carrier_delay,
          y = airlines,
          by = "carrier") |>
  select(carrier, name, avg_carr_dep_delay, avg_carr_arr_delay) |>
  arrange(avg_carr_dep_delay) |>
  head(5)
```

```
## # A tibble: 5 x 4
##   carrier name                  avg_carr_dep_delay avg_carr_arr_delay
##   <chr>   <chr>                              <dbl>              <dbl>
## 1 US      US Airways Inc.                     3.78               2.13
## 2 HA      Hawaiian Airlines Inc.              4.90              -6.92
```

```
## 3 AS      Alaska Airlines Inc.                5.80            -9.93
## 4 AA      American Airlines Inc.              8.59             0.364
## 5 DL      Delta Air Lines Inc.                9.26             1.64
```

```
inner_join(x = top5_carrier_speed, y = airlines, by = "carrier")
```

```
## # A tibble: 5 x 3
##   carrier avg_speed name
##   <chr>       <dbl> <chr>
## 1 HA           480. Hawaiian Airlines Inc.
## 2 VX           446. Virgin America
## 3 AS           444. Alaska Airlines Inc.
## 4 F9           425. Frontier Airlines Inc.
## 5 UA           421. United Air Lines Inc.
```

Filtering Join

```
anti_join(x = airlines,
          y = top5_carrier_speed,
          by = "carrier")
```

```
## # A tibble: 11 x 2
##    carrier name
##    <chr>   <chr>
##  1 9E      Endeavor Air Inc.
##  2 AA      American Airlines Inc.
##  3 B6      JetBlue Airways
##  4 DL      Delta Air Lines Inc.
##  5 EV      ExpressJet Airlines Inc.
##  6 FL      AirTran Airways Corporation
##  7 MQ      Envoy Air
##  8 OO      SkyWest Airlines Inc.
##  9 US      US Airways Inc.
## 10 WN      Southwest Airlines Co.
## 11 YV      Mesa Airlines Inc.
```

Exercises and Solutions

Question 1

Using the flights data set, which carrier have the highest average speed. Note: Remember to set na.rm = TRUE when calculating the average speed. 1. Hawaiian Airlines Inc. - ANSWER 2. Virgin America 3. Alaska Airlines Inc. 4. Frontier Airlines Inc. 5. United Air Lines Inc.

```
avg_speed_table <- flights |>
  mutate(speed = distance/air_time*60) |>
  group_by(carrier) |>
  summarize(avg_speed = mean(speed, na.rm = TRUE)) |>
```

```
  arrange(desc(avg_speed))

avg_speed_table
```

```
## # A tibble: 16 x 2
##    carrier avg_speed
##    <chr>       <dbl>
##  1 HA           480.
##  2 VX           446.
##  3 AS           444.
##  4 F9           425.
##  5 UA           421.
##  6 DL           418.
##  7 AA           417.
##  8 WN           401.
##  9 B6           400.
## 10 FL           394.
## 11 MQ           368.
## 12 OO           366.
## 13 EV           363.
## 14 9E           345.
## 15 US           342.
## 16 YV           332.
```

```
inner_join(x = avg_speed_table,
           y = airlines,
           by = "carrier")
```

```
## # A tibble: 16 x 3
##    carrier avg_speed name
##    <chr>       <dbl> <chr>
##  1 HA           480. Hawaiian Airlines Inc.
##  2 VX           446. Virgin America
##  3 AS           444. Alaska Airlines Inc.
##  4 F9           425. Frontier Airlines Inc.
##  5 UA           421. United Air Lines Inc.
##  6 DL           418. Delta Air Lines Inc.
##  7 AA           417. American Airlines Inc.
##  8 WN           401. Southwest Airlines Co.
##  9 B6           400. JetBlue Airways
## 10 FL           394. AirTran Airways Corporation
## 11 MQ           368. Envoy Air
## 12 OO           366. SkyWest Airlines Inc.
## 13 EV           363. ExpressJet Airlines Inc.
## 14 9E           345. Endeavor Air Inc.
## 15 US           342. US Airways Inc.
## 16 YV           332. Mesa Airlines Inc.
```

Question 2

How many flights in the month of December had no departure and arrrival delay. 1.39 2.37 3.32 4.31 ANSWER 5.34

```
flights |>
  filter(arr_delay == 0 & dep_delay == 0 & month == 12)
```

```
## # A tibble: 31 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    12     2     1030       1030       0    1155    1155       0 MQ
##  2  2013    12     2     1729       1729       0    2115    2115       0 VX
##  3  2013    12     4     1552       1552       0    1927    1927       0 DL
##  4  2013    12     7      759        759       0    1004    1004       0 B6
##  5  2013    12     7      949        949       0    1237    1237       0 B6
##  6  2013    12     7     1945       1945       0    2130    2130       0 MQ
##  7  2013    12     8      935        935       0    1115    1115       0 WN
##  8  2013    12     9      630        630       0     830     830       0 MQ
##  9  2013    12     9      945        945       0    1300    1300       0 AA
## 10  2013    12    11     1055       1055       0    1409    1409       0 DL
## # ... with 21 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

What is the distance covered in Kilometers for the flight with id number 4646 and tailnum N273WN. Note: A mile is 1.6 Kilometers 1. 115.625 Kilometers 2. 296 Kilometers - ANSWER 3. 290 Kilometers 4. 78 Kilometers 5. 234 Kilometers

```
flight_4646_N273WN <- flights |>
  filter(flight == 4646 & tailnum == "N273WN") |>
  mutate(dist_kil = distance*1.6)

flight_4646_N273WN |>
  select(dist_kil)
```

```
## # A tibble: 1 x 1
##   dist_kil
##      <dbl>
## 1      296
```

Question 4

The manufacturer of the plane in Question 3 is: 1. SIKORSKY 2. EMBRAER 3. AIRBUS 4. BOEING - ANSWER 5. GULFSTREAM AEROSPACE

```
inner_join(x = flight_4646_N273WN,
           y = planes,
           by = "tailnum") |>
  select(manufacturer)
```

```
## # A tibble: 1 x 1
##   manufacturer
##   <chr>
## 1 BOEING
```

Question 5

What is the tailnum of the fastest Air Plane departing New York. 1. N819AW 2. N382HA 3. N654UA 4. N228UA - ANSWER 5. N315AS

```
plane_speed <- flights |>
  mutate(speed = distance/(air_time/60)) |>
  select(tailnum, speed) |>
  group_by(tailnum) |>
  summarize(avg_speed = mean(speed, na.rm = TRUE)) |>
  arrange(desc(avg_speed)) |>
  head(5)

plane_speed
```

```
## # A tibble: 5 x 2
##   tailnum avg_speed
##   <chr>       <dbl>
## 1 N228UA       501.
## 2 N315AS       499.
## 3 N654UA       499.
## 4 N819AW       490.
## 5 N382HA       486.
```

```
inner_join(x = plane_speed,
           y = planes,
           by = "tailnum")
```

```
## # A tibble: 5 x 10
##   tailnum avg_speed  year type          manuf~1 model engines seats speed engine
##   <chr>       <dbl> <int> <chr>         <chr>   <chr>   <int> <int> <int> <chr>
## 1 N228UA       501.  2002 Fixed wing m~ BOEING  777-~       2   400    NA Turbo~
## 2 N315AS       499.  2002 Fixed wing m~ BOEING  737-~       2   149    NA Turbo~
## 3 N654UA       499.  1992 Fixed wing m~ BOEING  767-~       2   330    NA Turbo~
## 4 N819AW       490.  2000 Fixed wing m~ AIRBUS~ A319~       2   179    NA Turbo~
## 5 N382HA       486.  2010 Fixed wing m~ AIRBUS  A330~       2   377    NA Turbo~
## # ... with abbreviated variable name 1: manufacturer
```