# Joining Data with dplyr

Evans_G02504972

2023-10-31

First, lets load packages for this activity

```
library(tidyverse)
library(lubridate)
library(knitr)
library(skimr)
```

Next, we import the flights data

```
michiganFlights <- readRDS("fullMiFlights2021.rds")
```

```
view(michiganFlights)
```

```
list2env(michiganFlights, envir = .GlobalEnv)
```

```
## <environment: R_GlobalEnv>
```

Use the skim() and glimpse() functions to explore characteristics of some of the tables of data, setting the code chunk options to have include = FALSE, but echo = TRUE. Run the code chunk without knitting the document individually to explore patterns of missingness, variable names and types, etc.

```
glimpse(flights)
```

```
## Rows: 149,445
## Columns: 19
## $ year           <int> 2021, 2021, 2021, 2021, 2021, 2021, 2021, 2021, 2021, 2…
## $ month          <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…
## $ day            <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…
## $ dep_time       <int> 536, 557, 558, 600, 606, 610, 611, 611, 624, 624, 627, …
## $ sched_dep_time <int> 539, 600, 600, 607, 600, 615, 615, 616, 630, 615, 600, …
## $ dep_delay      <dbl> -3, -3, -2, -7, 6, -5, -4, -5, -6, 9, 27, -5, -5, -1, 0…
## $ arr_time       <int> 738, 758, 700, 820, 905, 809, 809, 804, 711, 806, 808, …
## $ sched_arr_time <int> 825, 748, 730, 831, 920, 832, 822, 826, 723, 800, 834, …
## $ arr_delay      <dbl> -47, 10, -30, -11, -15, -23, -13, -22, -12, 6, -26, -7,…
## $ carrier        <chr> "AA", "DL", "NK", "OH", "NK", "OH", "DL", "YX", "DL", "…
## $ flight         <int> 90, 174, 5, 512, 21, 507, 120, 491, 173, 284, 140, 157,…
## $ tailnum        <chr> "N750UW", "N354NB", "N653NK", "N507AE", "N675NK", "N600…
## $ origin         <chr> "DTW", "GRR", "DTW", "FNT", "DTW", "GRR", "DTW", "DTW",…
## $ dest           <chr> "PHX", "ATL", "LAS", "CLT", "FLL", "CLT", "ATL", "CLT",…
## $ air_time       <dbl> 227, 104, 220, 95, 162, 91, 95, 87, 28, 61, 207, 66, 15…
## $ distance       <dbl> 1671, 640, 1749, 555, 1127, 583, 594, 500, 120, 409, 15…
## $ hour           <dbl> 5, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 7, 7, 7, 7, 6, 7, 7, 7…
## $ minute         <dbl> 39, 0, 0, 7, 0, 15, 15, 16, 30, 15, 0, 0, 0, 0, 0, 50, …
## $ time_hour      <dttm> 2021-01-01 05:00:00, 2021-01-01 06:00:00, 2021-01-01 0…
```

```
glimpse(airports)
```

```
## Rows: 1,251
## Columns: 8
## $ faa   <chr> "AAF", "AAP", "ABE", "ABI", "ABL", "ABQ", "ABR", "ABY", "ACK", "…
## $ name  <chr> "Apalachicola Regional Airport", "Andrau Airpark", "Lehigh Valle…
## $ lat   <dbl> 29.72750, 29.72250, 40.65210, 32.41130, 67.10630, 35.04020, 45.4…
## $ lon   <dbl> -85.02750, -95.58830, -75.44080, -99.68190, -157.85699, -106.609…
## $ alt   <dbl> 20, 79, 393, 1791, 334, 5355, 1302, 197, 47, 516, 221, 75, 18, 7…
## $ tz    <dbl> -5, -6, -5, -6, -9, -7, -6, -5, -5, -6, -8, -5, -10, -6, -9, -6,…
## $ dst   <chr> "A", "A", "A", "A", "A", "A", "A", "A", "A", "A", "A", "A", "A",…
## $ tzone <chr> "America/New_York", "America/Chicago", "America/New_York", "Amer…
```

```
glimpse(weather)
```

```
## Rows: 34,897
## Columns: 15
## $ origin     <chr> "DTW", "DTW", "DTW", "DTW", "DTW", "DTW", "DTW", "DTW", "DT…
## $ year       <int> 2021, 2021, 2021, 2021, 2021, 2021, 2021, 2021, 2021, 2021,…
## $ month      <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,…
## $ day        <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,…
## $ hour       <int> 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 1…
## $ temp       <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,…
## $ dewp       <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,…
## $ humid      <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,…
## $ wind_dir   <dbl> 210, 0, 0, 0, 0, 0, 50, 90, 0, 0, 70, 60, 50, 60, 80, 70, 1…
## $ wind_speed <dbl> 5.75390, 0.00000, 0.00000, 0.00000, 0.00000, 0.00000, 2.301…
## $ wind_gust  <dbl> 6.621473, 0.000000, 0.000000, 0.000000, 0.000000, 0.000000,…
## $ precip     <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,…
## $ pressure   <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,…
## $ visib      <dbl> 8.0, 7.0, 6.0, 5.0, 4.0, 5.0, 5.0, 3.5, 3.0, 3.0, 3.0, 3.0,…
## $ time_hour  <dttm> 2021-01-01 00:00:00, 2021-01-01 01:00:00, 2021-01-01 02:00…
```

```
glimpse(planes)
```

```
## Rows: 3,962
## Columns: 9
## $ tailnum      <chr> "N101DQ", "N101DU", "N101HQ", "N102DN", "N102DU", "N102HQ…
## $ year         <int> 2020, 2018, 2007, 2020, NA, 2007, 1998, NA, 2020, 2007, 2…
## $ type         <chr> "Fixed wing multi engine", "Fixed wing multi engine", "Fi…
## $ manufacturer <chr> "AIRBUS", "C SERIES AIRCRAFT LTD PTNRSP", "EMBRAER-EMPRES…
## $ model        <chr> "A321-211", "BD-500-1A10", "ERJ 170-200 LR", "A321-211", …
## $ engines      <int> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, …
## $ seats        <int> 199, 133, 80, 199, 133, 80, 182, 133, 199, 80, 88, 182, 1…
## $ speed        <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …
## $ engine       <chr> "Turbo-fan", "Turbo-fan", "Turbo-fan", "Turbo-fan", "Turb…
```

```
skim(weather)
```

Data summary

| Name | weather |
|---|---|
| Number of rows | 34897 |
| Number of columns | 15 |
| _______________________ | |
| Column type frequency: | |
| character | 1 |
| numeric | 13 |
| POSIXct | 1 |
| ________________________ | |
| Group variables | None |

**Variable type: character**

| skim_variable | n_missing | complete_rate | min | max | empty | n_unique | whitespace |
|---|---|---|---|---|---|---|---|
| origin | 0 | 1 | 3 | 3 | 0 | 4 | 0 |

**Variable type: numeric**

| skim_variable | n_missing | complete_rate | mean | sd | p0 | p25 | p50 | p75 | p10 |
|---|---|---|---|---|---|---|---|---|---|
| year | 0 | 1.00 | 2021.00 | 0.00 | 2021.00 | 2021.00 | 2021.00 | 2021.00 | 202 |
| month | 0 | 1.00 | 6.51 | 3.44 | 1.00 | 4.00 | 7.00 | 9.00 | 12. |
| day | 0 | 1.00 | 15.67 | 8.77 | 1.00 | 8.00 | 16.00 | 23.00 | 31. |
| hour | 0 | 1.00 | 11.50 | 6.92 | 0.00 | 6.00 | 12.00 | 18.00 | 23. |
| temp | 34395 | 0.01 | 45.75 | 18.05 | 10.90 | 30.90 | 37.90 | 64.00 | 82. |
| dewp | 34396 | 0.01 | 36.89 | 18.27 | 5.00 | 21.90 | 30.00 | 51.10 | 73. |
| humid | 34397 | 0.01 | 72.73 | 15.76 | 31.52 | 61.59 | 72.07 | 87.06 | 100 |
| wind_dir | 952 | 0.97 | 181.80 | 107.41 | 0.00 | 90.00 | 200.00 | 270.00 | 360 |
| wind_speed | 465 | 0.99 | 8.09 | 5.36 | 0.00 | 4.60 | 8.06 | 11.51 | 36. |
| wind_gust | 465 | 0.99 | 9.31 | 6.17 | 0.00 | 5.30 | 9.27 | 13.24 | 42. |
| precip | 33636 | 0.04 | 0.01 | 0.03 | 0.00 | 0.00 | 0.01 | 0.01 | 0.4 |
| pressure | 34629 | 0.01 | 1011.31 | 7.07 | 1000.10 | 1004.38 | 1011.15 | 1018.50 | 102 |
| visib | 99 | 1.00 | 8.80 | 2.29 | 0.06 | 9.00 | 10.00 | 10.00 | 10. |

**Variable type: POSIXct**

| skim_variable | n_missing | complete_rate | min | max | median | n_unique |
|---|---|---|---|---|---|---|
| time_hour | 0 | 1 | 2021-01-01 | 2021-12-30 23:00:00 | 2021-07-01 23:00:00 | 8735 |

```
skim(flights)
```

Data summary

| | |
|---|---|
| Name | flights |
| Number of rows | 149445 |
| Number of columns | 19 |
| _______________________ | |
| Column type frequency: | |
| character | 4 |
| numeric | 14 |
| POSIXct | 1 |
| ________________________ | |
| Group variables | None |

**Variable type: character**

| skim_variable | n_missing | complete_rate | min | max | empty | n_unique | whitespace |
|---|---|---|---|---|---|---|---|
| carrier | 0 | 1 | 2 | 2 | 0 | 15 | 0 |
| tailnum | 117 | 1 | 5 | 6 | 0 | 4136 | 0 |

| skim_variable | n_missing | complete_rate | min | max | empty | n_unique | whitespace |
|---|---|---|---|---|---|---|---|
| origin | 0 | 1 | 3 | 3 | 0 | 4 | 0 |
| dest | 0 | 1 | 3 | 3 | 0 | 114 | 0 |

**Variable type: numeric**

| skim_variable | n_missing | complete_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
|---|---|---|---|---|---|---|---|---|---|---|
| year | 0 | 1.00 | 2021.00 | 0.00 | 2021 | 2021 | 2021 | 2021 | 2021 | ▁▁▇ |
| month | 0 | 1.00 | 6.77 | 3.35 | 1 | 4 | 7 | 10 | 12 | ▇▆▇ |
| day | 0 | 1.00 | 15.74 | 8.78 | 1 | 8 | 16 | 23 | 31 | ▇▇▇ |
| dep_time | 1382 | 0.99 | 1371.55 | 493.76 | 1 | 944 | 1353 | 1745 | 2400 | ▁▇▇ |
| sched_dep_time | 0 | 1.00 | 1369.27 | 486.06 | 500 | 930 | 1355 | 1737 | 2327 | ▇▇▆ |
| dep_delay | 1384 | 0.99 | 6.96 | 45.39 | -34 | -5 | -3 | 1 | 1948 | ▇▁▁ |
| arr_time | 1433 | 0.99 | 1464.41 | 517.39 | 1 | 1049 | 1455 | 1839 | 2400 | ▁▃▇ |
| sched_arr_time | 0 | 1.00 | 1481.47 | 507.65 | 1 | 1100 | 1504 | 1840 | 2359 | ▁▂▇ |
| arr_delay | 1715 | 0.99 | -0.22 | 47.48 | -79 | -17 | -9 | 1 | 1961 | ▇▁▁ |
| flight | 0 | 1.00 | 371.09 | 221.12 | 1 | 176 | 372 | 548 | 927 | ▇▇▇ |
| air_time | 1715 | 0.99 | 95.45 | 62.47 | 15 | 50 | 75 | 133 | 393 | ▇▃▁ |
| distance | 0 | 1.00 | 654.32 | 488.30 | 74 | 296 | 501 | 983 | 2986 | ▇▂▁ |
| hour | 0 | 1.00 | 13.42 | 4.83 | 5 | 9 | 13 | 17 | 23 | ▇▇▆ |
| minute | 0 | 1.00 | 27.52 | 18.71 | 0 | 10 | 29 | 45 | 59 | ▇▅▇ |

**Variable type: POSIXct**

| skim_variable | n_missing | complete_rate | min | max | median | n_unique |
|---|---|---|---|---|---|---|
| time_hour | 0 | 1 | 2021-01-01 05:00:00 | 2021-12-31 22:00:00 | 2021-07-12 21:00:00 | 6452 |

The flights and planes data set connect through which variable(s)?

the 'tailnum' variable connect the two table

The flights and airports data set connect through which variable(s)?

the 'faa','origin', and 'dest', connect the two tables #### The flights and weather data set connect through which variable(s)? the 'year','month', 'day ,'hour', and the location variable('origin') connect the two tables #### Suppose we wanted to draw (approximately) the route each plane flies from its origin to its destination. Which variables would we need? Which tables would we need to combine? We need flights dataset('origin' and 'dest') variables and airport dataset('faa','name'. we also need 'lat & 'lon' variables. #### Now suppose we wanted to explore typical weather patterns for departing flights at different airports and explore the weather's relationship with departure delays. Considering the wind speeds and amount of precipitation, which variables would we need for this? Which tables would we need to combine? the weather data set(all the linking variables and the 'wind speed' and 'precipitation' variables) and the flight data set(all the linking variables,'dep', 'delay').

## Outer Join

Combine the airlines and flights data frames with left_join() to create a new data set called flightsCarriers.

```
flightscarriers <- flights %>%
  left_join(airlines, by = c("carrier" = "carrier"))
flightscarriers
```

```
## # A tibble: 149,445 × 20
##     year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##    <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
##  1  2021     1     1      536            539        -3      738            825
##  2  2021     1     1      557            600        -3      758            748
##  3  2021     1     1      558            600        -2      700            730
##  4  2021     1     1      600            607        -7      820            831
##  5  2021     1     1      606            600         6      905            920
##  6  2021     1     1      610            615        -5      809            832
##  7  2021     1     1      611            615        -4      809            822
##  8  2021     1     1      611            616        -5      804            826
##  9  2021     1     1      624            630        -6      711            723
## 10  2021     1     1      624            615         9      806            800
## # ℹ 149,435 more rows
## # ℹ 12 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>, name <chr>
```

Calculate the average flight distance for each carrier using the full name of the carriers. Who had the longest flights on average? Who had the shortest?

```
#calculating avg flight distance for each carrier using the full name of the carrier. who ha
flightscarriers %>%
  group_by(name) %>%
  summarise(AvgDistance = mean(distance)) %>%
  arrange(AvgDistance)
```

```
## # A tibble: 15 × 2
##    name                      AvgDistance
##    <chr>                           <dbl>
##  1 Endeavor Air Inc.                328.
##  2 Envoy Air                        362.
##  3 Republic Airline                 384.
##  4 SkyWest Airlines Inc.            428.
##  5 PSA Airlines Inc.                510.
##  6 JetBlue Airways                  584.
##  7 Southwest Airlines Co.           600.
##  8 United Air Lines Inc.            679.
##  9 Mesa Airlines Inc.               764.
## 10 American Airlines Inc.           912.
## 11 Delta Air Lines Inc.             960.
## 12 Spirit Air Lines                1113.
## 13 Allegiant Air                   1117.
## 14 Frontier Airlines Inc.          1177.
## 15 Alaska Airlines Inc.            1927
```

Alaska airlines has the longest flight on average and endeovor air has the shortest

*Combine the flights and weather data frames with left_join() to create a new data set called flightsWeather. How many rows does flightsWeather have?*

```
#Add flight information to the weather data

weatherFlight <- weather %>%
  left_join(flights,
            by = c("origin", "year", "month", "day", "hour"))
weatherFlight
```

```
## # A tibble: 168,159 × 29
##    origin  year month   day  hour  temp  dewp humid wind_dir wind_speed
##    <chr>  <int> <int> <int> <dbl> <dbl> <dbl> <dbl>    <dbl>      <dbl>
##  1 DTW     2021     1     1     0    NA    NA    NA      210       5.75
##  2 DTW     2021     1     1     1    NA    NA    NA        0       0
##  3 DTW     2021     1     1     2    NA    NA    NA        0       0
##  4 DTW     2021     1     1     3    NA    NA    NA        0       0
##  5 DTW     2021     1     1     4    NA    NA    NA        0       0
##  6 DTW     2021     1     1     5    NA    NA    NA        0       0
##  7 DTW     2021     1     1     6    NA    NA    NA       50       2.30
##  8 DTW     2021     1     1     6    NA    NA    NA       50       2.30
##  9 DTW     2021     1     1     6    NA    NA    NA       50       2.30
## 10 DTW     2021     1     1     6    NA    NA    NA       50       2.30
## # ℹ 168,149 more rows
## # ℹ 19 more variables: wind_gust <dbl>, precip <dbl>, pressure <dbl>,
## #   visib <dbl>, time_hour.x <dttm>, dep_time <int>, sched_dep_time <int>,
## #   dep_delay <dbl>, arr_time <int>, sched_arr_time <int>, arr_delay <dbl>,
## #   carrier <chr>, flight <int>, tailnum <chr>, dest <chr>, air_time <dbl>,
## #   distance <dbl>, minute <dbl>, time_hour.y <dttm>
```

```
weatherFlight %>% nrow()
```

```
## [1] 168159
```

The ‘weatherflights’ data set has all the weather data, supplemented with flight data when available, and it has 168159 rows.This has more rows than the original ‘weather’ data set since some weather information was dublicated due to multiple flights occuring in the same hour at the same airport
Combine the weather and flights data frames with full_join() to create a new data set called weatherFlightsFull. How many rows does weatherFlightsFull have?

```
#Add weather  information to the weather data

flightsWeather <- flights %>%
  left_join(weather,
            by = c("origin", "year", "month", "day", "hour"))
```

The ‘flightsWeather’ data set has all the flights data, supplemented with weather data when available, and it has 149445 rows ### Combine the weather and flights data frames with full_join() to create a new data set called weatherFlightsFull. How many rows does weatherFlightsFull have?

```
#Add weather  information to the flight data

weatherFlightsFull <- flights %>%
  full_join(weather,
            by = c("origin", "year", "month", "day", "hour"))
weatherFlightsFull
```

```
## # A tibble: 168,504 × 29
##     year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##    <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
##  1  2021     1     1      536            539        -3      738            825
##  2  2021     1     1      557            600        -3      758            748
##  3  2021     1     1      558            600        -2      700            730
##  4  2021     1     1      600            607        -7      820            831
##  5  2021     1     1      606            600         6      905            920
##  6  2021     1     1      610            615        -5      809            832
##  7  2021     1     1      611            615        -4      809            822
##  8  2021     1     1      611            616        -5      804            826
##  9  2021     1     1      624            630        -6      711            723
## 10  2021     1     1      624            615         9      806            800
## # ℹ 168,494 more rows
## # ℹ 21 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.x <dttm>, temp <dbl>, dewp <dbl>,
## #   humid <dbl>, wind_dir <dbl>, wind_speed <dbl>, wind_gust <dbl>,
## #   precip <dbl>, pressure <dbl>, visib <dbl>, time_hour.y <dttm>
```

```
weatherFlightsFull %>% nrow()
```

```
## [1] 168504
```

the ‘weatherFlightsFull’ data set has all the flights data, supplemented with weather data when available, weather data even when no flights occured, and it has 168504 rows.

Since ‘weatherFullFlights’ has 168504 rows and weather flights has 168159 rows, there were 168504-168159 =345 flights with no weather information available.

## Considering all of the data we have available, how many flights have missing wind speeds?

```
#Using flightsWeather to answer this question

flightsWeather %>%
  dplyr::pull(wind_speed) %>%
  is.na() %>%
  sum()
```

```
## [1] 1526
```

There were 1526 flights that had missing wind speeds. # Inner joins

### Combine the weather and flights data frames with inner_join() to create a new data set called innerWeatherFlights. How many rows does innerWeatherFlights have?

```
#Add weather  information to the flight data

innerweatherFlightsFull <- flights %>%
  inner_join(weather,
             by = c("origin", "year", "month", "day", "hour"))
```

```
innerweatherFlightsFull %>% nrow()
```

```
## [1] 149100
```

The ‘innerweatherFlightsFull’ data set has information on flights that had weather information available, it has 149100 rows, there were 149445-149100=345 flights with no weather information available