Combining Multiple Tables

Today we will start to look at the nycflights13 data tables.

You can think of this dataset as if it was a spreadsheet with multiple spreadsheets within a workbook.

Take a look at the *flights* table and the *airlines* table. Do the two tables have a common variable that can be used as a key to match the rows of the tables?

```
library(tidyverse)
library(nycflights13)
library(skimr)

flights %>% head()
```

```
## # A tibble: 6 x 19
##    year month   day dep_time sched_dep_time dep_delay arr_time
##   <int> <int> <int>    <int>          <int>     <dbl>    <int>
## 1  2013     1     1      517            515         2      830
## 2  2013     1     1      533            529         4      850
## 3  2013     1     1      542            540         2      923
## 4  2013     1     1      544            545        -1     1004
## 5  2013     1     1      554            600        -6      812
## 6  2013     1     1      554            558        -4      740
## # ... with 12 more variables: sched_arr_time <int>, 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>
```

```
airlines %>% head()
```

```
## # 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.
```

```
airports %>% head()
```

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

```
flightsJoined <- flights %>%
  inner_join(airlines, by = c("carrier" = "carrier") )
```

```
flightsJoined %>% head()
```

```
## # A tibble: 6 x 20
##    year month   day dep_time sched_dep_time dep_delay arr_time
##   <int> <int> <int>    <int>          <int>     <dbl>    <int>
## 1  2013     1     1      517            515         2      830
## 2  2013     1     1      533            529         4      850
## 3  2013     1     1      542            540         2      923
## 4  2013     1     1      544            545        -1     1004
## 5  2013     1     1      554            600        -6      812
## 6  2013     1     1      554            558        -4      740
## # ... with 13 more variables: sched_arr_time <int>, 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>
```

```
glimpse(flightsJoined)
```

```
## Observations: 336,776
## Variables: 20
## $ year           <int> 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013,...
## $ month          <int> 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,...
## $ dep_time       <int> 517, 533, 542, 544, 554, 554, 555, 557, 557, 55...
## $ sched_dep_time <int> 515, 529, 540, 545, 600, 558, 600, 600, 600, 60...
## $ dep_delay      <dbl> 2, 4, 2, -1, -6, -4, -5, -3, -3, -2, -2, -2, -2...
## $ arr_time       <int> 830, 850, 923, 1004, 812, 740, 913, 709, 838, 7...
## $ sched_arr_time <int> 819, 830, 850, 1022, 837, 728, 854, 723, 846, 7...
## $ arr_delay      <dbl> 11, 20, 33, -18, -25, 12, 19, -14, -8, 8, -2, -...
## $ carrier        <chr> "UA", "UA", "AA", "B6", "DL", "UA", "B6", "EV",...
## $ flight         <int> 1545, 1714, 1141, 725, 461, 1696, 507, 5708, 79...
## $ tailnum        <chr> "N14228", "N24211", "N619AA", "N804JB", "N668DN...
## $ origin         <chr> "EWR", "LGA", "JFK", "JFK", "LGA", "EWR", "EWR"...
## $ dest           <chr> "IAH", "IAH", "MIA", "BQN", "ATL", "ORD", "FLL"...
## $ air_time       <dbl> 227, 227, 160, 183, 116, 150, 158, 53, 140, 138...
## $ distance       <dbl> 1400, 1416, 1089, 1576, 762, 719, 1065, 229, 94...
## $ hour           <dbl> 5, 5, 5, 5, 6, 5, 6, 6, 6, 6, 6, 6, 6, 6, 6, 5,...
## $ minute         <dbl> 15, 29, 40, 45, 0, 58, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ time_hour      <dttm> 2013-01-01 05:00:00, 2013-01-01 05:00:00, 2013...
## $ name           <chr> "United Air Lines Inc.", "United Air Lines Inc....
```

A package that I like for summarizing the variables in a dataframe is the *skim* function from the *skimr* package. What do you think?

Check that the new column *name* has been added.

```
flightsJoined %>% select(carrier, name, flight, origin, dest) %>%
  head()
```

```
## # A tibble: 6 x 5
##   carrier name                  flight origin dest
##   <chr>   <chr>                  <int> <chr>  <chr>
## 1 UA      United Air Lines Inc.   1545 EWR    IAH
## 2 UA      United Air Lines Inc.   1714 LGA    IAH
## 3 AA      American Airlines Inc.  1141 JFK    MIA
## 4 B6      JetBlue Airways          725 JFK    BQN
```

```
## 5 DL      Delta Air Lines Inc.      461 LGA    ATL
## 6 UA      United Air Lines Inc.    1696 EWR    ORD
```

How big are the dataframes? Note the base R finction *nrow* just give a number as output, using the dplyr function *count()* give a dataframe.

```
nrow(flights)
```

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

```
nrow(flightsJoined)
```

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

```
flights %>% count()
```

```
## # A tibble: 1 x 1
##        n
##    <int>
## 1 336776
```

```
flightsJoined %>% count()
```

```
## # A tibble: 1 x 1
##        n
##    <int>
## 1 336776
```

Suppose we are interested in the flights from NYC airports to the West Coast.

The Pacific Time Zone is time zone -8 in the airports table. (I do not know why there are more airports that given in the book.)

```
airportsPT <- airports %>% filter(tz == -8)

airportsPT %>% head()
```

```
## # A tibble: 6 x 8
##   faa   name                  lat   lon   alt    tz dst   tzone
##   <chr> <chr>               <dbl> <dbl> <int> <dbl> <chr> <chr>
## 1 0S9   Jefferson County I~  48.1 -123.   108    -8 A     America/Los_Ang~
## 2 1C9   Frazier Lake Airpa~  54.0 -125.   152    -8 A     America/Vancouv~
## 3 1RL   Point Roberts Airp~  49.0 -123.    10    -8 A     America/Los_Ang~
## 4 38W   Lynden Airport       49.0 -122.   106    -8 A     America/Los_Ang~
## 5 49X   Chemehuevi Valley    34.5 -114.   638    -8 A     America/Los_Ang~
## 6 55S   Packwood             46.4 -121.  1057    -8 A     America/Los_Ang~
```

```
airportsPT %>% count()
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1   178
```

Now if we are interested in the flights from NYC to the west coast, find the airports in the Pacific Time Zone and join the airportPT we will get the flights to the west coast.

```
nycDestPT <- flights %>% inner_join(airportsPT, by = c("dest" = "faa") )

nycDestPT %>% count()
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1 46324
```

If we *left_join* we will get all of the rows in *flights*.

Check out the *map* function, it applies a function to all columns.

```
nycDest <- flights %>% left_join(airportsPT, by = c("dest" = "faa") )

nycDest %>% count()
```

```
## # A tibble: 1 x 1
##        n
##    <int>
## 1 336776
```

```
nycDest %>% head()
```

```
## # A tibble: 6 x 26
##    year month   day dep_time sched_dep_time dep_delay arr_time
##   <int> <int> <int>    <int>          <int>     <dbl>    <int>
## 1  2013     1     1      517            515         2      830
## 2  2013     1     1      533            529         4      850
## 3  2013     1     1      542            540         2      923
## 4  2013     1     1      544            545        -1     1004
## 5  2013     1     1      554            600        -6      812
## 6  2013     1     1      554            558        -4      740
## # ... with 19 more variables: sched_arr_time <int>, 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>, lat <dbl>, lon <dbl>, alt <int>,
## #   tz <dbl>, dst <chr>, tzone <chr>
```

```
nycDest %>% map(~sum(is.na(.)))
```

```
## $year
## [1] 0
##
## $month
## [1] 0
##
## $day
## [1] 0
##
## $dep_time
## [1] 8255
##
## $sched_dep_time
## [1] 0
##
## $dep_delay
## [1] 8255
##
## $arr_time
## [1] 8713
```

```
##
## $sched_arr_time
## [1] 0
##
## $arr_delay
## [1] 9430
##
## $carrier
## [1] 0
##
## $flight
## [1] 0
##
## $tailnum
## [1] 2512
##
## $origin
## [1] 0
##
## $dest
## [1] 0
##
## $air_time
## [1] 9430
##
## $distance
## [1] 0
##
## $hour
## [1] 0
##
## $minute
## [1] 0
##
## $time_hour
## [1] 0
##
## $name
## [1] 290452
##
## $lat
## [1] 290452
##
## $lon
## [1] 290452
##
## $alt
## [1] 290452
##
## $tz
## [1] 290452
##
## $dst
## [1] 290452
```

```
##
## $tzone
## [1] 290452
```