Lecture Assignment 14

Viraj Vijaywargiya

2022-05-18

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(nycflights13)
```

Part 13.2.1

Question 1

To draw (approximately) the route each plane flies from its origin to its destination, we require the longitude and latitude of the origin and destination airports, therefore, we need the **origin**, **dest**, **lon**, and **lat** variables. For this, we would need to combine the **flights** table**,** which has the origin and dest variable, and the **airports** table, which has the lon and lat variable.

Question 2

weather connects to airports through **origin** (variable in the weather table) and **faa** (variable in the airports table). In the diagram, it can be represented with an arrow going from **faa** in airports to **origin** in weather.

Question 3

If weather contained records for all airports in the USA, other than just the origin (NYC) airports, then it would also have the weather for the destination airports. Therefore, flights would additionally connect with weather via **dest**.

Question 4

The data frame would be a table having variables representing special dates, like **year**, **month**, **day**, and **holiday**. The primary keys would be year, month, and day columns. Below in an example of how the data frame might look like,

```r
special <- tribble(
  ~year, ~month, ~day, ~holiday,
  #----/--/--/----
  2013, 01, 01, "New Years Day",
  2013, 01, 21, "Martin Luther King Jr. Day",
  2013, 02, 18, "Presidents' Day",
  2013, 07, 04, "Independence Day",
  2013, 11, 28, "Thanksgiving Day",
  2013, 12, 25, "Christmas Day"
)
```

This table above would connect to the flights table through the **year**, **month**, and **day** variables.

Part 13.3.1

Question 1

```
flights %>%
  mutate(num_flight = row_number()) %>%
  glimpse()
```

```
## Rows: 336,776
## Columns: 20
## $ year           <int> 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 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> 517, 533, 542, 544, 554, 554, 555, 557, 557, 558, 558, ~
## $ sched_dep_time <int> 515, 529, 540, 545, 600, 558, 600, 600, 600, 600, 600, ~
## $ dep_delay      <dbl> 2, 4, 2, -1, -6, -4, -5, -3, -3, -2, -2, -2, -2, -2, -1~
## $ arr_time       <int> 830, 850, 923, 1004, 812, 740, 913, 709, 838, 753, 849,~
## $ sched_arr_time <int> 819, 830, 850, 1022, 837, 728, 854, 723, 846, 745, 851,~
## $ arr_delay      <dbl> 11, 20, 33, -18, -25, 12, 19, -14, -8, 8, -2, -3, 7, -1~
## $ carrier        <chr> "UA", "UA", "AA", "B6", "DL", "UA", "B6", "EV", "B6", "~
## $ flight         <int> 1545, 1714, 1141, 725, 461, 1696, 507, 5708, 79, 301, 4~
## $ tailnum        <chr> "N14228", "N24211", "N619AA", "N804JB", "N668DN", "N394~
## $ origin         <chr> "EWR", "LGA", "JFK", "JFK", "LGA", "EWR", "EWR", "LGA",~
## $ dest           <chr> "IAH", "IAH", "MIA", "BQN", "ATL", "ORD", "FLL", "IAD",~
## $ air_time       <dbl> 227, 227, 160, 183, 116, 150, 158, 53, 140, 138, 149, 1~
## $ distance       <dbl> 1400, 1416, 1089, 1576, 762, 719, 1065, 229, 944, 733, ~
## $ hour           <dbl> 5, 5, 5, 5, 6, 5, 6, 6, 6, 6, 6, 6, 6, 6, 6, 5, 6, 6, 6~
## $ minute         <dbl> 15, 29, 40, 45, 0, 58, 0, 0, 0, 0, 0, 0, 0, 0, 0, 59, 0~
## $ time_hour      <dttm> 2013-01-01 05:00:00, 2013-01-01 05:00:00, 2013-01-01 0~
## $ num_flight     <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, ~
```

I added the column, num_flight, as a surrogate key.

Question 2

```
glimpse(Lahman::Batting)
```

```
## Rows: 108,789
## Columns: 22
## $ playerID <chr> "abercda01", "addybo01", "allisar01", "allisdo01", "ansonca01~
## $ yearID   <int> 1871, 1871, 1871, 1871, 1871, 1871, 1871, 1871, 1871, 1871, 1~
## $ stint    <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
## $ teamID   <fct> TRO, RC1, CL1, WS3, RC1, FW1, RC1, BS1, FW1, BS1, CL1, CL1, W~
## $ lgID     <fct> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ G        <int> 1, 25, 29, 27, 25, 12, 1, 31, 1, 18, 22, 1, 10, 3, 20, 29, 1,~
## $ AB       <int> 4, 118, 137, 133, 120, 49, 4, 157, 5, 86, 89, 3, 36, 15, 94, ~
## $ R        <int> 0, 30, 28, 28, 29, 9, 0, 66, 1, 13, 18, 0, 6, 7, 24, 26, 0, 0~
## $ H        <int> 0, 32, 40, 44, 39, 11, 1, 63, 1, 13, 27, 0, 7, 6, 33, 32, 0, ~
## $ X2B      <int> 0, 6, 4, 10, 11, 2, 0, 10, 1, 2, 1, 0, 0, 0, 9, 3, 0, 0, 1, 0~
## $ X3B      <int> 0, 0, 5, 2, 3, 1, 0, 9, 0, 1, 10, 0, 0, 0, 1, 3, 0, 0, 1, 0, ~
## $ HR       <int> 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 3, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0~
## $ RBI      <int> 0, 13, 19, 27, 16, 5, 2, 34, 1, 11, 18, 0, 1, 5, 21, 23, 0, 0~
## $ SB       <int> 0, 8, 3, 1, 6, 0, 0, 11, 0, 1, 0, 0, 2, 2, 4, 4, 0, 0, 3, 0, ~
## $ CS       <int> 0, 1, 1, 1, 2, 1, 0, 6, 0, 0, 1, 0, 0, 0, 0, 4, 0, 0, 1, 0, 0~
## $ BB       <int> 0, 4, 2, 0, 2, 0, 1, 13, 0, 0, 3, 1, 2, 0, 2, 9, 0, 0, 4, 1, ~
## $ SO       <int> 0, 0, 5, 2, 1, 1, 0, 1, 0, 0, 4, 0, 0, 0, 2, 2, 3, 0, 2, 0, 2~
## $ IBB      <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ HBP      <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ SH       <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ SF       <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ GIDP     <int> 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 2, 0, 1, 2, 0, 0, 0, 0, 3~
```

The primary key for the above dataset (Lahman::Batting) is **playerID**, **yearID**, and **stint**.

```
library(babynames)
glimpse(babynames::babynames)
```

```
## Rows: 1,924,665
## Columns: 5
## $ year <dbl> 1880, 1880, 1880, 1880, 1880, 1880, 1880, 1880, 1880, 1880, 1880,~
## $ sex  <chr> "F", "F", "F", "F", "F", "F", "F", "F", "F", "F", "F", "F", "F", ~
## $ name <chr> "Mary", "Anna", "Emma", "Elizabeth", "Minnie", "Margaret", "Ida",~
## $ n    <int> 7065, 2604, 2003, 1939, 1746, 1578, 1472, 1414, 1320, 1288, 1258,~
## $ prop <dbl> 0.07238359, 0.02667896, 0.02052149, 0.01986579, 0.01788843, 0.016~
```

The primary key for the above dataset (babynames::babynames) is year, sex, and name.

```
library(nasaweather)
```

```
##
## Attaching package: 'nasaweather'

## The following object is masked from 'package:dplyr':
##
##     storms
```

```
glimpse(nasaweather::atmos)
```

```
## Rows: 41,472
## Columns: 11
## $ lat       <dbl> 36.200000, 33.704348, 31.208696, 28.713043, 26.217391, 23.72~
## $ long      <dbl> -113.8000, -113.8000, -113.8000, -113.8000, -113.8000, -113.~
## $ year      <int> 1995, 1995, 1995, 1995, 1995, 1995, 1995, 1995, 1995, 1995, ~
## $ month     <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ surftemp  <dbl> 272.7, 279.5, 284.7, 289.3, 292.2, 294.1, 295.0, 298.3, 300.~
## $ temp      <dbl> 272.1, 282.2, 285.2, 290.7, 292.7, 293.6, 294.6, 296.9, 297.~
## $ pressure  <dbl> 835, 940, 960, 990, 1000, 1000, 1000, 1000, 1000, 1000, 1000~
## $ ozone     <dbl> 304, 304, 298, 276, 274, 264, 258, 252, 250, 250, 248, 248, ~
## $ cloudlow  <dbl> 7.5, 11.5, 16.5, 20.5, 26.0, 30.0, 29.5, 26.5, 27.5, 26.0, 2~
## $ cloudmid  <dbl> 34.5, 32.5, 26.0, 14.5, 10.5, 9.5, 11.0, 17.5, 18.5, 16.5, 1~
## $ cloudhigh <dbl> 26.0, 20.0, 16.0, 13.0, 7.5, 8.0, 14.5, 19.5, 22.5, 21.0, 19~
```

The primary key for the above dataset(nasaweather::atmos) is lat, long, year, and month.

```
library(fueleconomy)
glimpse(fueleconomy::vehicles)
```

```
## Rows: 33,442
## Columns: 12
## $ id    <dbl> 13309, 13310, 13311, 14038, 14039, 14040, 14834, 14835, 14836, 1~
## $ make  <chr> "Acura", "Acura", "Acura", "Acura", "Acura", "Acura", "Acura", "~
## $ model <chr> "2.2CL/3.0CL", "2.2CL/3.0CL", "2.2CL/3.0CL", "2.3CL/3.0CL", "2.3~
## $ year  <dbl> 1997, 1997, 1997, 1998, 1998, 1998, 1999, 1999, 1999, 1995, 1996~
## $ class <chr> "Subcompact Cars", "Subcompact Cars", "Subcompact Cars", "Subcom~
## $ trans <chr> "Automatic 4-spd", "Manual 5-spd", "Automatic 4-spd", "Automatic~
## $ drive <chr> "Front-Wheel Drive", "Front-Wheel Drive", "Front-Wheel Drive", "~
## $ cyl   <dbl> 4, 4, 6, 4, 4, 6, 4, 4, 6, 5, 5, 6, 5, 6, 5, 6, 6, 6, 6, 6, 6, 6~
## $ displ <dbl> 2.2, 2.2, 3.0, 2.3, 2.3, 3.0, 2.3, 2.3, 3.0, 2.5, 2.5, 3.2, 2.5,~
## $ fuel  <chr> "Regular", "Regular", "Regular", "Regular", "Regular", "Regular"~
## $ hwy   <dbl> 26, 28, 26, 27, 29, 26, 27, 29, 26, 23, 23, 22, 23, 22, 23, 22, ~
## $ cty   <dbl> 20, 22, 18, 19, 21, 17, 20, 21, 17, 18, 18, 17, 18, 17, 17, 17, ~
```

The primary key for the above dataset (fueleconomy::vehicles) is id.

```
glimpse(diamonds)
```

```
## Rows: 53,940
## Columns: 10
## $ carat   <dbl> 0.23, 0.21, 0.23, 0.29, 0.31, 0.24, 0.24, 0.26, 0.22, 0.23, 0.~
## $ cut     <ord> Ideal, Premium, Good, Premium, Good, Very Good, Very Good, Ver~
## $ color   <ord> E, E, E, I, J, J, I, H, E, H, J, J, F, J, E, E, I, J, J, J, I,~
## $ clarity <ord> SI2, SI1, VS1, VS2, SI2, VVS2, VVS1, SI1, VS2, VS1, SI1, VS1, ~
## $ depth   <dbl> 61.5, 59.8, 56.9, 62.4, 63.3, 62.8, 62.3, 61.9, 65.1, 59.4, 64~
## $ table   <dbl> 55, 61, 65, 58, 58, 57, 57, 55, 61, 61, 55, 56, 61, 54, 62, 58~
## $ price   <int> 326, 326, 327, 334, 335, 336, 336, 337, 337, 338, 339, 340, 34~
## $ x       <dbl> 3.95, 3.89, 4.05, 4.20, 4.34, 3.94, 3.95, 4.07, 3.87, 4.00, 4.~
## $ y       <dbl> 3.98, 3.84, 4.07, 4.23, 4.35, 3.96, 3.98, 4.11, 3.78, 4.05, 4.~
## $ z       <dbl> 2.43, 2.31, 2.31, 2.63, 2.75, 2.48, 2.47, 2.53, 2.49, 2.39, 2.~
```

There is no combination of variables/columns that uniquely identifies the each observation, therefore, there is no primary key for the above dataset (fueleconomy::vehicles).