# Lab 4

Brandon Leslie

## Lab 4 - Transforming data, cont: mutating joins

### Toy data-sets

**Create the toy data sets in the slides using tibble rather than tribble**

#### Band d.s.

```
band <- tibble(
 name = c("Mick", "John", "Paul"),
 band = c("Stones", "Beatles", "Beatles")
)

band
```

```
# A tibble: 3 x 2
  name  band
  <chr> <chr>
1 Mick  Stones
2 John  Beatles
3 Paul  Beatles
```

#### Instruments d.s.

```r
instrument <- tibble(
 name = c("John", "Paul", "Keith"),
 band = c("Guitar", "Bass", "Guitar")
)

instrument
```

```
# A tibble: 3 x 2
  name  band
  <chr> <chr>
1 John  Guitar
2 Paul  Bass
3 Keith Guitar
```

**Experiment with the various joins as in the examples**

### Left Join

```r
band %>%
  left_join(instrument, by = "name")
```

```
# A tibble: 3 x 3
  name  band.x  band.y
  <chr> <chr>   <chr>
1 Mick  Stones  <NA>
2 John  Beatles Guitar
3 Paul  Beatles Bass
```

### Right join

```r
band %>%
  right_join(instrument, by = "name")
```

```
# A tibble: 3 x 3
  name  band.x  band.y
  <chr> <chr>   <chr>
1 John  Beatles Guitar
2 Paul  Beatles Bass
3 Keith <NA>    Guitar
```

### Full Join

```
band %>%
  full_join(instrument, by = "name")
```

```
# A tibble: 4 x 3
  name  band.x  band.y
  <chr> <chr>   <chr>
1 Mick  Stones  <NA>
2 John  Beatles Guitar
3 Paul  Beatles Bass
4 Keith <NA>    Guitar
```

### Inner

```
band %>%
  inner_join(instrument, by = "name")
```

```
# A tibble: 2 x 3
  name  band.x  band.y
  <chr> <chr>   <chr>
1 John  Beatles Guitar
2 Paul  Beatles Bass
```

## NYC Flights Data-set

**Use data(package = "nycflights13") to see the included data sets.**

```
data(package = "nycflights13")
```

**Use skimr package to inspect data frames**

```
skim(flights)
```

Table 1: Data summary

| | |
|---|---|
| Name | flights |
| Number of rows | 336776 |
| 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 | 1.00 | 2 | 2 | 0 | 16 | 0 |
| tailnum | 2512 | 0.99 | 5 | 6 | 0 | 4043 | 0 |
| origin | 0 | 1.00 | 3 | 3 | 0 | 3 | 0 |
| dest | 0 | 1.00 | 3 | 3 | 0 | 105 | 0 |

**Variable type: numeric**

| skim_variable | n_missing | complete_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
|---|---|---|---|---|---|---|---|---|---|---|
| year | 0 | 1.00 | 2013.00 | 0.00 | 2013 | 2013 | 2013 | 2013 | 2013 | |
| month | 0 | 1.00 | 6.55 | 3.41 | 1 | 4 | 7 | 10 | 12 | |
| day | 0 | 1.00 | 15.71 | 8.77 | 1 | 8 | 16 | 23 | 31 | |
| dep_time | 8255 | 0.98 | 1349.11 | 488.28 | 1 | 907 | 1401 | 1744 | 2400 | |
| sched_dep_time | 0 | 1.00 | 1344.25 | 467.34 | 106 | 906 | 1359 | 1729 | 2359 | |
| dep_delay | 8255 | 0.98 | 12.64 | 40.21 | -43 | -5 | -2 | 11 | 1301 | |
| arr_time | 8713 | 0.97 | 1502.05 | 533.26 | 1 | 1104 | 1535 | 1940 | 2400 | |
| sched_arr_time | 0 | 1.00 | 1536.38 | 497.46 | 1 | 1124 | 1556 | 1945 | 2359 | |
| arr_delay | 9430 | 0.97 | 6.90 | 44.63 | -86 | -17 | -5 | 14 | 1272 | |
| flight | 0 | 1.00 | 1971.92 | 1632.47 | 1 | 553 | 1496 | 3465 | 8500 | |
| air_time | 9430 | 0.97 | 150.69 | 93.69 | 20 | 82 | 129 | 192 | 695 | |
| distance | 0 | 1.00 | 1039.91 | 733.23 | 17 | 502 | 872 | 1389 | 4983 | |
| hour | 0 | 1.00 | 13.18 | 4.66 | 1 | 9 | 13 | 17 | 23 | |
| minute | 0 | 1.00 | 26.23 | 19.30 | 0 | 8 | 29 | 44 | 59 | |

**Variable type: POSIXct**

| skim_variable | n_missing | complete_rate | min | max | median | n_unique |
|---|---|---|---|---|---|---|
| time_hour | 0 | 1 | 2013-01-01 05:00:00 | 2013-12-31 23:00:00 | 2013-07-03 10:00:00 | 6936 |

```
skim(airlines)
```

Table 5: Data summary

| | |
|---|---|
| Name | airlines |
| Number of rows | 16 |
| Number of columns | 2 |
| _____ | |
| Column type frequency: | |
| character | 2 |
| _____ | |
| Group variables | None |

**Variable type: character**

| skim_variable | n_missing | complete_rate | min | max | empty | n_unique | whitespace |
|---|---|---|---|---|---|---|---|
| carrier | 0 | 1 | 2 | 2 | 0 | 16 | 0 |
| name | 0 | 1 | 9 | 27 | 0 | 16 | 0 |

```
skim(weather)
```

Table 7: Data summary

| | |
|---|---|
| Name | weather |
| Number of rows | 26115 |
| 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 | 3 | 0 |

**Variable type: numeric**

| skim_variable | n_missing | complete_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
|---|---|---|---|---|---|---|---|---|---|---|
| year | 0 | 1.00 | 2013.00 | 0.00 | 2013.00 | 2013.00 | 2013.00 | 2013.00 | 2013.00 | |
| month | 0 | 1.00 | 6.50 | 3.44 | 1.00 | 4.00 | 7.00 | 9.00 | 12.00 | |
| day | 0 | 1.00 | 15.68 | 8.76 | 1.00 | 8.00 | 16.00 | 23.00 | 31.00 | |
| hour | 0 | 1.00 | 11.49 | 6.91 | 0.00 | 6.00 | 11.00 | 17.00 | 23.00 | |
| temp | 1 | 1.00 | 55.26 | 17.79 | 10.94 | 39.92 | 55.40 | 69.98 | 100.04 | |
| dewp | 1 | 1.00 | 41.44 | 19.39 | -9.94 | 26.06 | 42.08 | 57.92 | 78.08 | |
| humid | 1 | 1.00 | 62.53 | 19.40 | 12.74 | 47.05 | 61.79 | 78.79 | 100.00 | |
| wind_dir | 460 | 0.98 | 199.76 | 107.31 | 0.00 | 120.00 | 220.00 | 290.00 | 360.00 | |
| wind_speed | 4 | 1.00 | 10.52 | 8.54 | 0.00 | 6.90 | 10.36 | 13.81 | 1048.36 | |
| wind_gust | 20778 | 0.20 | 25.49 | 5.95 | 16.11 | 20.71 | 24.17 | 28.77 | 66.75 | |
| precip | 0 | 1.00 | 0.00 | 0.03 | 0.00 | 0.00 | 0.00 | 0.00 | 1.21 | |
| pressure | 2729 | 0.90 | 1017.90 | 7.42 | 983.80 | 1012.90 | 1017.60 | 1023.00 | 1042.10 | |
| visib | 0 | 1.00 | 9.26 | 2.06 | 0.00 | 10.00 | 10.00 | 10.00 | 10.00 | |

**Variable type: POSIXct**

| skim_variable | n_missing | complete_rate | min | max | median | n_unique |
|---|---|---|---|---|---|---|
| time_hour | 0 | 1 | 2013-01-01 01:00:00 | 2013-12-30 18:00:00 | 2013-07-01 14:00:00 | 8714 |

```
skim(planes)
```

Table 11: Data summary

| | |
|---|---|
| Name | planes |
| Number of rows | 3322 |
| Number of columns | 9 |

| | |
|---|---|
| Column type frequency: | |
| character | 5 |
| numeric | 4 |
| Group variables | None |

**Variable type: character**

| skim_variable | n_missing | complete_rate | min | max | empty | n_unique | whitespace |
|---|---|---|---|---|---|---|---|
| tailnum | 0 | 1 | 5 | 6 | 0 | 3322 | 0 |
| type | 0 | 1 | 10 | 24 | 0 | 3 | 0 |
| manufacturer | 0 | 1 | 4 | 29 | 0 | 35 | 0 |
| model | 0 | 1 | 2 | 18 | 0 | 127 | 0 |
| engine | 0 | 1 | 7 | 13 | 0 | 6 | 0 |

**Variable type: numeric**

| skim_variable | n_missing | complete_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
|---|---|---|---|---|---|---|---|---|---|---|
| year | 70 | 0.98 | 2000.48 | 7.19 | 1956 | 1997.0 | 2001 | 2005 | 2013 | |
| engines | 0 | 1.00 | 2.00 | 0.12 | 1 | 2.0 | 2 | 2 | 4 | |
| seats | 0 | 1.00 | 154.32 | 73.65 | 2 | 140.0 | 149 | 182 | 450 | |
| speed | 3299 | 0.01 | 236.78 | 149.76 | 90 | 107.5 | 162 | 432 | 432 | |

```
skim(airports)
```

Table 14: Data summary

| | |
|---|---|
| Name | airports |
| Number of rows | 1458 |
| Number of columns | 8 |
| Column type frequency: | |
| character | 4 |
| numeric | 4 |
| Group variables | None |

**Variable type: character**

| skim_variable | n_missing | complete_rate | min | max | empty | n_unique | whitespace |
|---|---|---|---|---|---|---|---|
| faa | 0 | 1 | 3 | 3 | 0 | 1458 | 0 |
| name | 0 | 1 | 4 | 51 | 0 | 1440 | 0 |
| dst | 0 | 1 | 1 | 1 | 0 | 3 | 0 |
| tzone | 3 | 1 | 14 | 19 | 0 | 9 | 0 |

**Variable type: numeric**

| skim_variable | n_missing | complete_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
|---|---|---|---|---|---|---|---|---|---|---|
| lat | 0 | 1 | 41.65 | 10.45 | 19.72 | 34.26 | 40.09 | 45.07 | 72.27 | |
| lon | 0 | 1 | -103.39 | 29.84 | -176.65 | -119.19 | -94.66 | -82.52 | 174.11 | |
| alt | 0 | 1 | 1001.42 | 1523.63 | -54.00 | 70.25 | 473.00 | 1062.50 | 9078.00 | |
| tz | 0 | 1 | -6.52 | 1.62 | -10.00 | -8.00 | -6.00 | -5.00 | 8.00 | |

**Join airlines to flights to find the names of which airlines had the largest arrival delays.**

```
flights %>%
  left_join(airlines, join_by(carrier)) %>%
  select(name, arr_delay, tailnum) %>%
  arrange(desc(arr_delay))
```

```
# A tibble: 336,776 x 3
   name                    arr_delay tailnum
   <chr>                       <dbl> <chr>
 1 Hawaiian Airlines Inc.       1272 N384HA
 2 Envoy Air                    1127 N504MQ
 3 Envoy Air                    1109 N517MQ
 4 American Airlines Inc.       1007 N338AA
 5 Envoy Air                     989 N665MQ
 6 Delta Air Lines Inc.          931 N959DL
 7 Delta Air Lines Inc.          915 N927DA
 8 Delta Air Lines Inc.          895 N6716C
 9 American Airlines Inc.        878 N5DMAA
```

```
10 Envoy Air                       875 N523MQ
# i 336,766 more rows
```

The names of the airlines with the largest **arr_delay**'s are: **Hawaiian Airlines Inc.**, **Envoy Air**, **American Airlines Inc.**, and **Delta Air Lines Inc.**

**Use flights and airports to compute the distance and average arr_delay by destination airport (names only, not codes).**

```
flights2.0 <- flights %>%
 left_join(airports, join_by(dest == name)) %>%
  group_by("Destination Airport" = dest) %>%
  summarize(
    "Average Arrival Delay" = mean(arr_delay, na.rm = TRUE),
    "Distance" = sum(distance, na.rm = TRUE)

  )
flights2.0
```

```
# A tibble: 105 x 3
   `Destination Airport` `Average Arrival Delay` Distance
   <chr>                                   <dbl>    <dbl>
 1 ABQ                                      4.38   463804
 2 ACK                                      4.85    52735
 3 ALB                                     14.4     62777
 4 ANC                                     -2.5     26960
 5 ATL                                     11.3  13033618
 6 AUS                                      6.02  3693263
 7 AVL                                      8.00   160485
 8 BDL                                      7.05    51388
 9 BGR                                      8.03   141750
10 BHM                                     16.9    257201
# i 95 more rows
```

### Population and Continents data sets, Brightspace

**Join these two data sets, and create a new data frame called population_summary that contains a row for each continent and a column for the total population for that continent, in descending order of population.**

```r
population <- read_csv("data/population.csv")
continents <- read_csv("data/continents.csv")

population_summary  <- population %>%
 left_join(continents,  join_by(country == entity)) %>%
  group_by("Continents" = continent) %>%
  summarize(
    Total_Population = sum(population, na.rm = TRUE)
  ) %>%
  arrange(desc(Total_Population))
population_summary
```

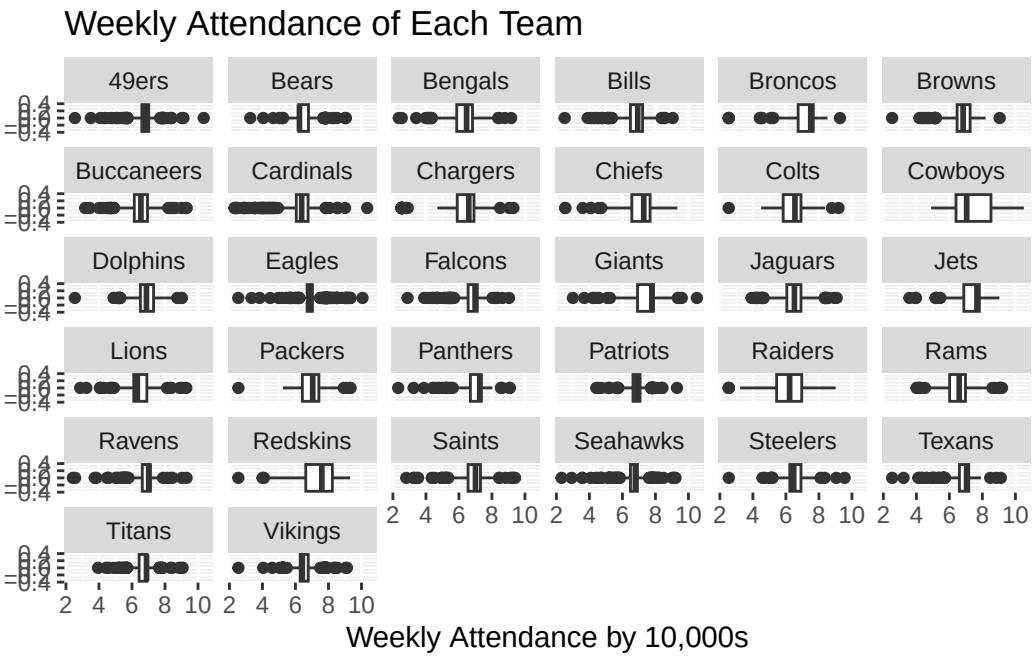
```
# A tibble: 6 x 2
  Continents    Total_Population
  <chr>                    <dbl>
1 Asia                   4685922.
2 Africa                 1424831.
3 Europe                  740607.
4 North America           594415.
5 South America           437233.
6 Oceania                  44752.
```

### NFL `attendance` and `standings` data sets, Brightspace

**Make boxplots of weekly attendance for each team**

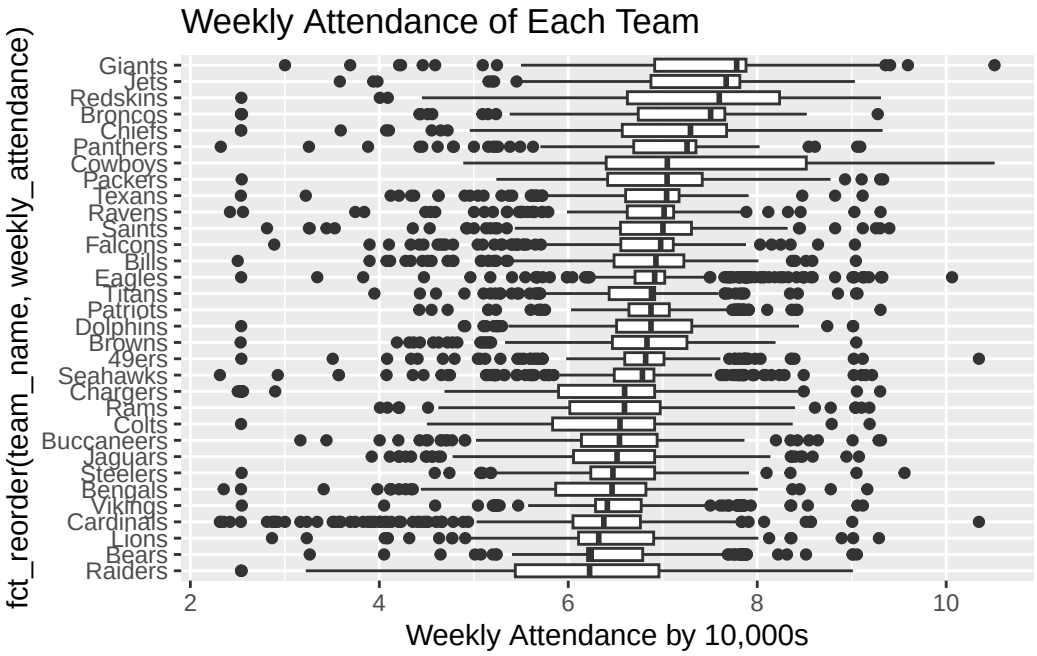
```r
attendance <- read_csv("data/attendance.csv")


attendance %>%
  ggplot(aes(x = weekly_attendance)) +
  geom_boxplot() +
    labs( x = "Weekly Attendance by 10,000s",
          title = "Weekly Attendance of Each Team") +
    scale_x_continuous(label = label_number(scale = 1/10000)) +
  facet_wrap(~team_name)
```




**Use `fct_reorder` to order the teams according to attendance**

```
attendance %>%
  ggplot(aes(y = fct_reorder(team_name, weekly_attendance), x = weekly_attendance)) +
  geom_boxplot() +
    labs( x = "Weekly Attendance by 10,000s",
          title = "Weekly Attendance of Each Team") +
    scale_x_continuous(label = label_number(scale = 1/10000))
```



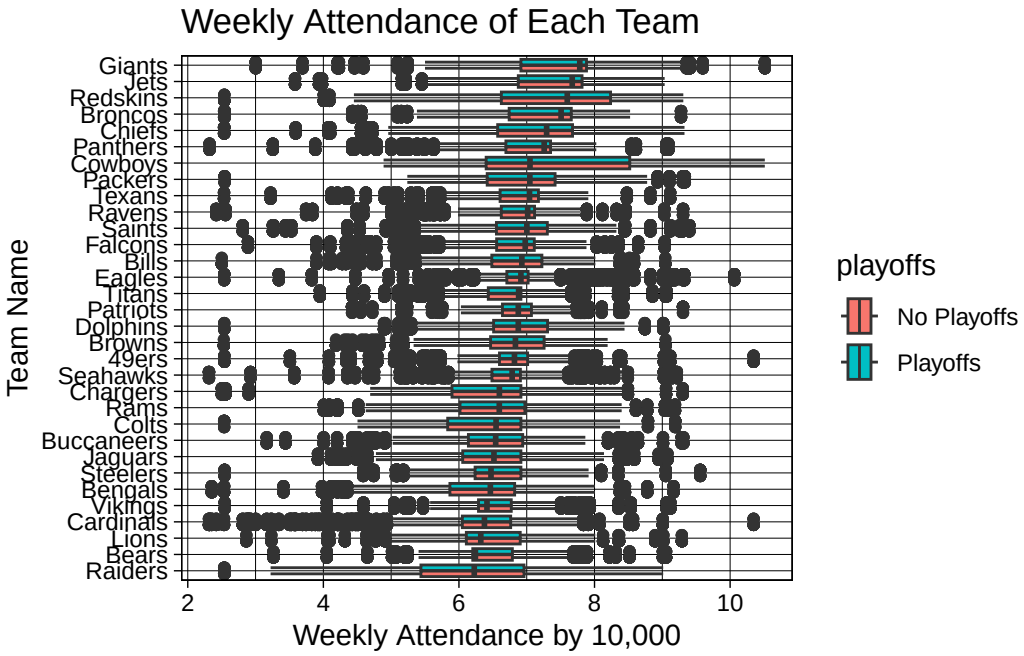

**Does it make sense to arrange vertically or horizontally?**

It makes sense to arrange vertically because it is more visually appealing to the eye to compare this way.

**Use `fill =` to plot side by side for playoffs vs not**

```
standings <- read_csv("data/standings.csv")

attendance %>%
  left_join(standings, join_by(team_name)) %>%
  ggplot(aes(y = fct_reorder(team_name, weekly_attendance), x = weekly_attendance)) +
  geom_boxplot(aes(fill = playoffs)) +
    labs( x = "Weekly Attendance by 10,000",
          y = "Team Name",
          title = "Weekly Attendance of Each Team") +
    scale_x_continuous(label = label_number(scale = 1/10000)) +
  theme_linedraw()
```


Weekly Attendance of Each Team
Team Name
Giants
Jets
Redskins
Broncos
Chiefs
Panthers
Cowboys
Packers
Texans
Ravens
Saints
Falcons
Bills
Eagles
Titans
Patriots
Dolphins
Browns
49ers
Seahawks
Chargers
Rams
Colts
Buccaneers
Jaguars
Steelers
Bengals
Vikings
Cardinals
Lions
Bears
Raiders
2
4
6
8
10
Weekly Attendance by 10,000
playoffs
No Playoffs
Playoffs