# TIDYING DATA

UNIVERSITY OF
CALGARY

# TIDYING DATA IS STRUCTURING DATASETS TO FACILITATE ANALYSIS

[Hadley Wickham 2014]

“Happy families (and tidy data) are all alike; every unhappy family (and messy dataset) is unhappy in its own way.”

[Leo Tolstory and Hadley Wickham 2014]

# PRINCIPLES OF TIDY DATA

Each variable forms a column

Each observation forms a row

Each type of observational unit forms a table

# UNTIDY DATASET EXAMPLE

Words spoken in the Lord of the Rings Trilogy

**The Fellowship of the Ring**

| Race | Female | Male |
|---|---|---|
| Elf | 1229 | 971 |
| Hobbit | 14 | 3644 |
| Man | 0 | 1995 |

**The Two Towers**

| Race | Female | Male |
|---|---|---|
| Elf | 331 | 513 |
| Hobbit | 0 | 2463 |
| Man | 401 | 3589 |

**The Return of the King**

| Race | Female | Male |
|---|---|---|
| Elf | 183 | 510 |
| Hobbit | 2 | 2673 |
| Man | 268 | 2459 |

How many variables are there in this dataset?

[Example from Jenny Bryan]

# TIDIED DATASET EXAMPLE

Each variable forms a column

Each observation forms a row

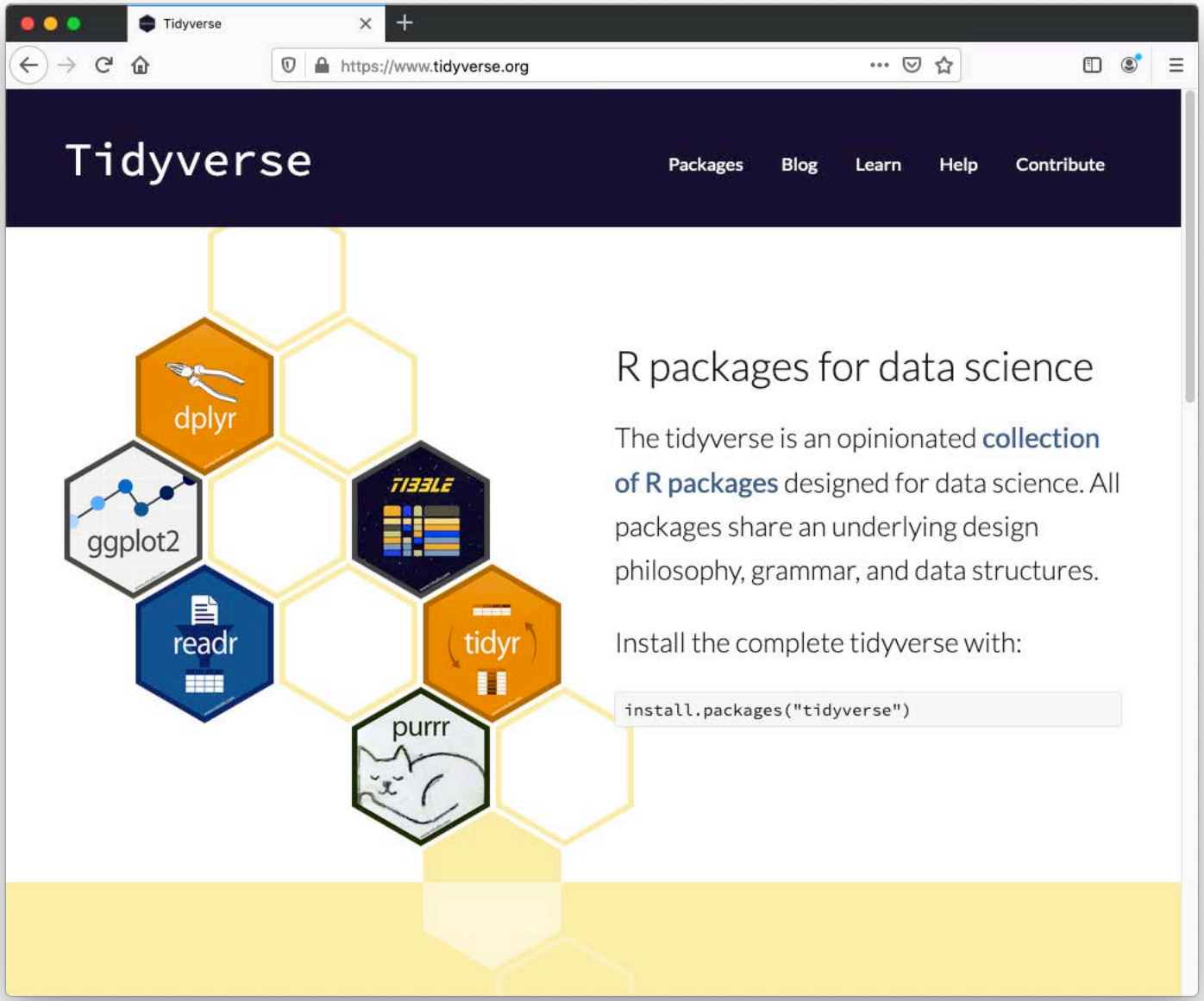
Each type of observational unit forms a table

| Film | Gender | Race | Words |
|---|---|---|---|
| The Fellowship Of The Ring | Female | Elf | 1229 |
| The Fellowship Of The Ring | Male | Elf | 971 |
| The Fellowship Of The Ring | Female | Hobbit | 14 |
| The Fellowship Of The Ring | Male | Hobbit | 3644 |
| The Fellowship Of The Ring | Female | Man | 0 |
| The Fellowship Of The Ring | Male | Man | 1995 |
| The Two Towers | Female | Elf | 331 |
| The Two Towers | Male | Elf | 513 |
| The Two Towers | Female | Hobbit | 0 |
| The Two Towers | Male | Hobbit | 2463 |
| The Two Towers | Female | Man | 401 |
| The Two Towers | Male | Man | 3589 |
| The Return Of The King | Female | Elf | 183 |
| The Return Of The King | Male | Elf | 510 |
| The Return Of The King | Female | Hobbit | 2 |
| The Return Of The King | Male | Hobbit | 2673 |
| The Return Of The King | Female | Man | 268 |
| The Return Of The King | Male | Man | 2459 |

WHEN STARTING A NEW ANALYSIS, GET DATA INTO A TIDY FORMAT FIRST

# PART OF THE FOUNDING PRINCIPLES BEHIND THE OF THE TIDYVERSE PACKAGES FOR R (DPLYR, READR, ETC.)

# BUT EQUALLY RELEVANT IN OTHER ECOSYSTEMS



# COMMON CAUSES OF MESSY DATA

Column headers are values, not variable names

Multiple variables are stored in one column

Variables are stored in both rows and columns

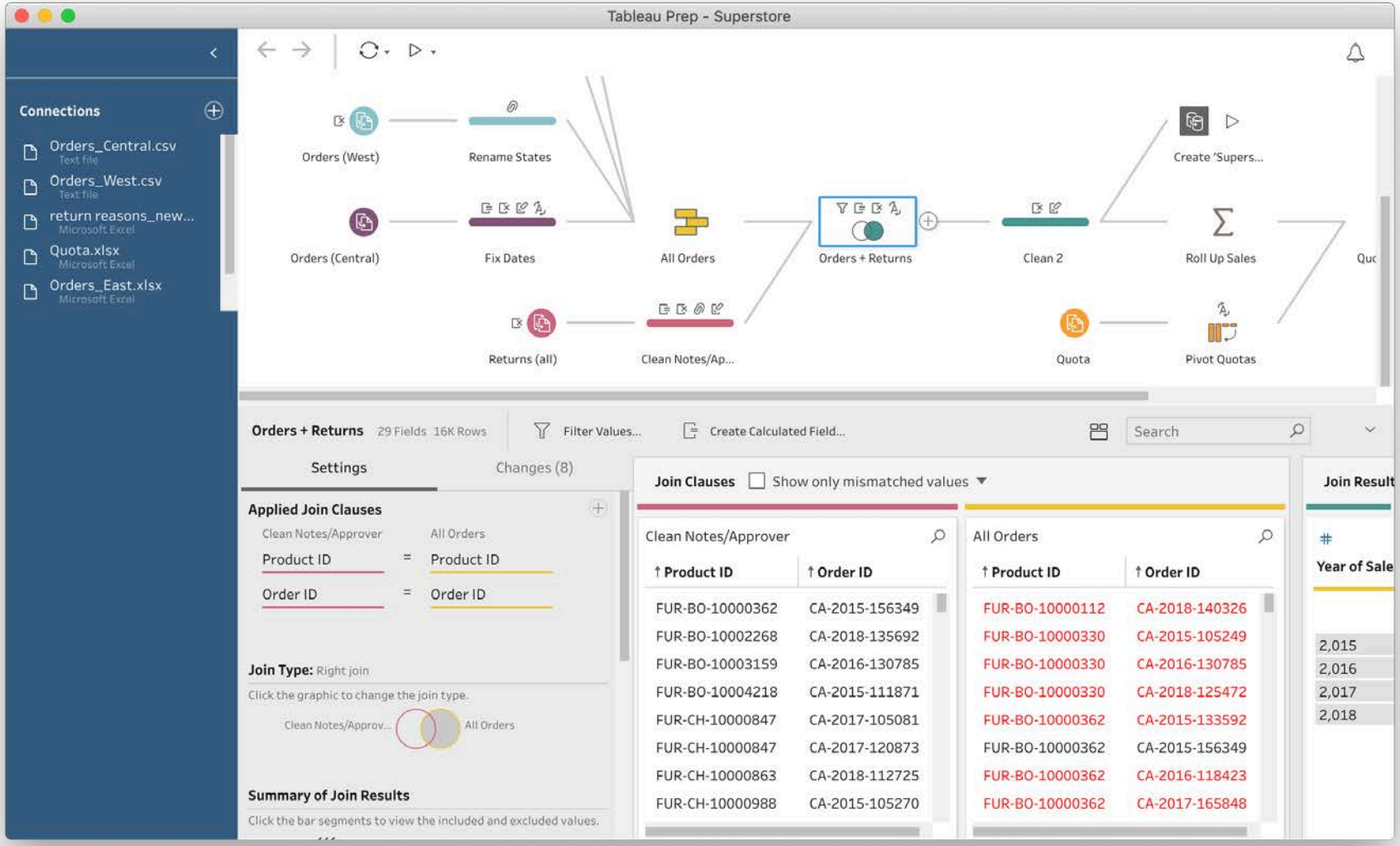
Multiple types of observational units are stored in the same table

A single observational unit is stored in multiple tables

# SOME DATA TIDYING DEMOS

## In Tableau Prep

(but you can tidy and wrangle equally well in lots of other tools)



# PROBLEM: COLUMN HEADERS ARE VALUES

## Billboard top hits of 2000

| | year | artist.inverted | track | time | genre | date.entered | date.peaked | x1st.week | x2nd.week | x3rd.week | ... | x67th.week |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 246 | 2000 | 2 Pac | Baby Don't Cry (Keep Ya Head Up II) | 4:22 | Rap | 2000-02-26 | 2000-03-11 | 87 | 82.0 | 72.0 | ... | NaN |
| 287 | 2000 | 2Ge+her | The Hardest Part Of Breaking Up (Is Getting Ba... | 3:15 | R&B | 2000-09-02 | 2000-09-09 | 91 | 87.0 | 92.0 | ... | NaN |
| 24 | 2000 | 3 Doors Down | Kryptonite | 3:53 | Rock | 2000-04-08 | 2000-11-11 | 81 | 70.0 | 68.0 | ... | NaN |
| 193 | 2000 | 3 Doors Down | Loser | 4:24 | Rock | 2000-10-21 | 2000-12-02 | 76 | 76.0 | 72.0 | ... | NaN |
| 69 | 2000 | 504 Boyz | Wobble Wobble | 3:35 | Rap | 2000-04-15 | 2000-05-06 | 57 | 34.0 | 25.0 | ... | NaN |
| 22 | 2000 | 98¡ | Give Me Just One Night (Una Noche) | 3:24 | Rock | 2000-08-19 | 2000-09-30 | 51 | 39.0 | 34.0 | ... | NaN |
| 304 | 2000 | A*Teens | Dancing Queen | 3:44 | Pop | 2000-07-08 | 2000-07-29 | 97 | 97.0 | 96.0 | ... | NaN |
| 14 | 2000 | Aaliyah | Try Again | 4:03 | Rock | 2000-03-18 | 2000-06-17 | 59 | 53.0 | 38.0 | ... | NaN |

These columns have song ranking each week after its initial appearance on the top 100 (if the song has dropped off the top 100, the element is empty).

These column headers are values of the variable “week,” and should be integrated into one column.

# PROBLEM: COLUMN HEADERS ARE VALUES

## Melting turns columns into rows

```
pd.melt(df, id_vars=['A'], value_vars=['B', 'C'])
```

| | A | B | C |
|---|---|---|---|
| 0 | a | 1 | 2 |
| 1 | b | 3 | 4 |
| 2 | c | 5 | 6 |

▶

| | A | variable | value |
|---|---|---|---|
| 0 | a | B | 1 |
| 1 | b | B | 3 |
| 2 | c | B | 5 |
| 3 | a | C | 2 |
| 4 | b | C | 4 |
| 5 | c | C | 6 |

| | year | artist.inverted | track | time | genre | date.entered | date.peaked | week | ranking |
|---|---|---|---|---|---|---|---|---|---|
| 246 | 2000 | 2 Pac | Baby Don't Cry (Keep Ya Head Up II) | 4:22 | Rap | 2000-02-26 | 2000-03-11 | 1 | 87.0 |
| 563 | 2000 | 2 Pac | Baby Don't Cry (Keep Ya Head Up II) | 4:22 | Rap | 2000-02-26 | 2000-03-11 | 2 | 82.0 |
| 880 | 2000 | 2 Pac | Baby Don't Cry (Keep Ya Head Up II) | 4:22 | Rap | 2000-02-26 | 2000-03-11 | 3 | 72.0 |
| 1197 | 2000 | 2 Pac | Baby Don't Cry (Keep Ya Head Up II) | 4:22 | Rap | 2000-02-26 | 2000-03-11 | 4 | 77.0 |
| 1514 | 2000 | 2 Pac | Baby Don't Cry (Keep Ya Head Up II) | 4:22 | Rap | 2000-02-26 | 2000-03-11 | 5 | 87.0 |
| 1831 | 2000 | 2 Pac | Baby Don't Cry (Keep Ya Head Up II) | 4:22 | Rap | 2000-02-26 | 2000-03-11 | 6 | 94.0 |
| 2148 | 2000 | 2 Pac | Baby Don't Cry (Keep Ya Head Up II) | 4:22 | Rap | 2000-02-26 | 2000-03-11 | 7 | 99.0 |
| 287 | 2000 | 2Ge+her | The Hardest Part Of Breaking Up (Is Getting Ba... | 3:15 | R&B | 2000-09-02 | 2000-09-09 | 1 | 91.0 |
| 604 | 2000 | 2Ge+her | The Hardest Part Of Breaking Up (Is Getting Ba... | 3:15 | R&B | 2000-09-02 | 2000-09-09 | 2 | 87.0 |
| 921 | 2000 | 2Ge+her | The Hardest Part Of Breaking Up (Is Getting Ba... | 3:15 | R&B | 2000-09-02 | 2000-09-09 | 3 | 92.0 |
| 24 | 2000 | 3 Doors Down | Kryptonite | 3:53 | Rock | 2000-04-08 | 2000-11-11 | 1 | 81.0 |
| 341 | 2000 | 3 Doors Down | Kryptonite | 3:53 | Rock | 2000-04-08 | 2000-11-11 | 2 | 70.0 |
| 658 | 2000 | 3 Doors Down | Kryptonite | 3:53 | Rock | 2000-04-08 | 2000-11-11 | 3 | 68.0 |
| 975 | 2000 | 3 Doors Down | Kryptonite | 3:53 | Rock | 2000-04-08 | 2000-11-11 | 4 | 67.0 |
| 1292 | 2000 | 3 Doors Down | Kryptonite | 3:53 | Rock | 2000-04-08 | 2000-11-11 | 5 | 66.0 |

# PROBLEM: MULTIPLE VARIABLES IN ONE COLUMN

## Tuberculosis Cases

| | country | year | column | cases |
|---|---|---|---|---|
| 0 | AD | 2000 | m014 | 0 |
| 1 | AD | 2000 | m1524 | 0 |
| 2 | AD | 2000 | m2534 | 1 |
| 3 | AD | 2000 | m3544 | 0 |
| 4 | AD | 2000 | m4554 | 0 |
| 5 | AD | 2000 | m5564 | 0 |
| 6 | AD | 2000 | m65 | 0 |
| 7 | AE | 2000 | m014 | 2 |
| 8 | AE | 2000 | m1524 | 4 |
| 9 | AE | 2000 | m2534 | 4 |
| 10 | AE | 2000 | m3544 | 6 |
| 11 | AE | 2000 | m4554 | 5 |
| 12 | AE | 2000 | m5564 | 12 |
| 13 | AE | 2000 | m65 | 10 |
| 14 | AE | 2000 | f014 | 3 |

The column "column" should be separated into separate gender and age columns

| | country | year | gender | age | cases |
|---|---|---|---|---|---|
| 0 | AD | 2000 | m | 0-14 | 0 |
| 1 | AD | 2000 | m | 15-24 | 0 |
| 2 | AD | 2000 | m | 25-34 | 1 |
| 3 | AD | 2000 | m | 35-44 | 0 |
| 4 | AD | 2000 | m | 45-54 | 0 |
| 5 | AD | 2000 | m | 55-64 | 0 |
| 6 | AD | 2000 | m | 65+ | 0 |
| 7 | AE | 2000 | m | 0-14 | 2 |
| 8 | AE | 2000 | m | 15-24 | 4 |
| 9 | AE | 2000 | m | 25-34 | 4 |
| 10 | AE | 2000 | m | 35-44 | 6 |
| 11 | AE | 2000 | m | 45-54 | 5 |
| 12 | AE | 2000 | m | 55-64 | 12 |
| 13 | AE | 2000 | m | 65+ | 10 |
| 14 | AE | 2000 | f | 0-14 | 3 |

# PROBLEM: VARIABLES STORED IN ROWS AND COLUMS

Weather Dataset

This column should be separated into two separate columns for tmax and tmin (the max and min temperature)

| | id | year | month | element | d1 | d2 | d3 |
|---|---|---|---|---|---|---|---|
| 0 | MX17004 | 2010 | 1 | tmax | NaN | NaN | NaN |
| 1 | MX17004 | 2010 | 1 | tmin | NaN | NaN | NaN |
| 2 | MX17004 | 2010 | 2 | tmax | NaN | 27.3 | 24.1 |
| 3 | MX17004 | 2010 | 2 | tmin | NaN | 14.4 | 14.4 |
| 4 | MX17004 | 2010 | 3 | tmax | NaN | NaN | NaN |
| 5 | MX17004 | 2010 | 3 | tmin | NaN | NaN | NaN |
| 6 | MX17004 | 2010 | 4 | tmax | NaN | NaN | NaN |
| 7 | MX17004 | 2010 | 4 | tmin | NaN | NaN | NaN |
| 8 | MX17004 | 2010 | 5 | tmax | NaN | NaN | NaN |
| 9 | MX17004 | 2010 | 5 | tmin | NaN | NaN | NaN |
| 10 | MX17004 | 2010 | 6 | tmax | NaN | NaN | NaN |

These are temperature values for a given day of the month and should be integrated into one value column.

# PROBLEM: VARIABLES STORED IN ROWS AND COLUMS

| | id | element | value | date |
|---|---|---|---|---|
| 638 | MX17004 | tmax | 27.8 | 2010-1-30 |
| 639 | MX17004 | tmin | 14.5 | 2010-1-30 |
| 24 | MX17004 | tmax | 27.3 | 2010-2-2 |
| 25 | MX17004 | tmin | 14.4 | 2010-2-2 |
| 46 | MX17004 | tmax | 24.1 | 2010-2-3 |
| 47 | MX17004 | tmin | 14.4 | 2010-2-3 |
| 222 | MX17004 | tmax | 29.7 | 2010-2-11 |
| 223 | MX17004 | tmin | 13.4 | 2010-2-11 |
| 486 | MX17004 | tmax | 29.9 | 2010-2-23 |
| 487 | MX17004 | tmin | 10.7 | 2010-2-23 |
| 92 | MX17004 | tmax | 32.1 | 2010-3-5 |

First, melt the multiple days columns into one "day" column.

Then add "year," "month," and "day" to one date column

# PROBLEM: VARIABLES STORED IN ROWS AND COLUMS

## Pivoting turns unique columns into rows

df.pivot(index='foo', columns='bar', values='baz')

| | foo | bar | baz | zoo |
|---|---|---|---|---|
| 0 | one | A | 1 | x |
| 1 | one | B | 2 | y |
| 2 | one | C | 3 | z |
| 3 | two | A | 4 | q |
| 4 | two | B | 5 | w |
| 5 | two | C | 6 | t |

| | A | B | C |
|---|---|---|---|
| one | 1 | 2 | 3 |
| two | 4 | 5 | 6 |

| element | id | date | tmax | tmin |
|---|---|---|---|---|
| 0 | MX17004 | 2010-1-30 | 27.8 | 14.5 |
| 1 | MX17004 | 2010-10-14 | 29.5 | 13.0 |
| 2 | MX17004 | 2010-10-15 | 28.7 | 10.5 |
| 3 | MX17004 | 2010-10-28 | 31.2 | 15.0 |
| 4 | MX17004 | 2010-10-5 | 27.0 | 14.0 |
| 5 | MX17004 | 2010-10-7 | 28.1 | 12.9 |
| 6 | MX17004 | 2010-11-2 | 31.3 | 16.3 |
| 7 | MX17004 | 2010-11-26 | 28.1 | 12.1 |
| 8 | MX17004 | 2010-11-27 | 27.7 | 14.2 |
| 9 | MX17004 | 2010-11-4 | 27.2 | 12.0 |
| 10 | MX17004 | 2010-11-5 | 26.3 | 7.9 |

# PROBLEM: MULTIPLE OBSERVATIONAL UNITS STORED IN THE SAME TABLE

## Songs

| | id | year | artist.inverted | track | time | genre | date.entered | date.peaked |
|---|---|---|---|---|---|---|---|---|
| 246 | 0 | 2000 | 2 Pac | Baby Don't Cry (Keep Ya Head Up II) | 4:22 | Rap | 2000-02-26 | 2000-03-11 |
| 287 | 1 | 2000 | 2Ge+her | The Hardest Part Of Breaking Up (Is Getting Ba... | 3:15 | R&B | 2000-09-02 | 2000-09-09 |
| 24 | 2 | 2000 | 3 Doors Down | Kryptonite | 3:53 | Rock | 2000-04-08 | 2000-11-11 |
| 193 | 3 | 2000 | 3 Doors Down | Loser | 4:24 | Rock | 2000-10-21 | 2000-12-02 |
| 69 | 4 | 2000 | 504 Boyz | Wobble Wobble | 3:35 | Rap | 2000-04-15 | 2000-05-06 |
| 22 | 5 | 2000 | 98¡ | Give Me Just One Night (Una Noche) | 3:24 | Rock | 2000-08-19 | 2000-09-30 |
| 304 | 6 | 2000 | A*Teens | Dancing Queen | 3:44 | Pop | 2000-07-08 | 2000-07-29 |
| 135 | 7 | 2000 | Aaliyah | I Don't Wanna | 4:15 | Rock | 2000-01-29 | 2000-03-04 |
| 14 | 8 | 2000 | Aaliyah | Try Again | 4:03 | Rock | 2000-03-18 | 2000-06-17 |
| 200 | 9 | 2000 | Adams, Yolanda | Open My Heart | 5:30 | Gospel | 2000-08-26 | 2000-10-21 |
| 227 | 10 | 2000 | Adkins, Trace | More | 3:05 | Country | 2000-04-29 | 2000-06-17 |
| 4 | 11 | 2000 | Aguilera, Christina | Come On Over Baby (All I Want Is You) | 3:38 | Rock | 2000-08-05 | 2000-10-14 |
| 26 | 12 | 2000 | Aguilera, Christina | I Turn To You | 4:00 | Rock | 2000-04-15 | 2000-07-01 |
| 11 | 13 | 2000 | Aguilera, Christina | What A Girl Wants | 3:18 | Rock | 1999-11-27 | 2000-01-15 |
| 109 | 14 | 2000 | Alice Deejay | Better Off Alone | 6:50 | Electronica | 2000-04-08 | 2000-06-03 |

The Billboard dataset could be separated into songs and ranking tables to remove unnecessary duplication and to check for errors

## Ranking

| | id | week | ranking |
|---|---|---|---|
| 246 | 0 | 1 | 87.0 |
| 563 | 0 | 2 | 82.0 |
| 874 | 0 | 3 | 72.0 |
| 1181 | 0 | 4 | 77.0 |
| 1481 | 0 | 5 | 87.0 |
| 1766 | 0 | 6 | 94.0 |
| 2043 | 0 | 7 | 99.0 |
| 287 | 1 | 1 | 91.0 |
| 603 | 1 | 2 | 87.0 |
| 914 | 1 | 3 | 92.0 |
| 24 | 2 | 1 | 81.0 |
| 341 | 2 | 2 | 70.0 |
| 653 | 2 | 3 | 68.0 |
| 960 | 2 | 4 | 67.0 |
| 1260 | 2 | 5 | 66.0 |

# PROBLEM: SINGLE OBSERVATIONAL UNIT STORED IN MULTIPLE TABLES

1. Read the files into a list of tables.

2. For each table, add a new column that records the original file name (because the file name is often the value of an important variable).

3. Combine all tables into a single table.