# Data Normalization

- **Introduction to Data Management**
  - Some definitions to get started and for future reference
    - Note: terminology varies across texts, platforms and applications – concept is the same
  - Database basics

- **Database Design**
  - Use database specifications/real-world models to identify tables, columns and keys for a database
  - (E)ER diagrams, translating ER diagrams to relational DB schema

- **Data Normalization**
  - Rules of data normalization
  - Steps for normalizing database schema
  - Concepts of “tidy” data, transforming raw data to processed data

# Data Normalization

- The first three normal forms

| Normal form | Description |
|---|---|
| First (1NF) | The value stored at the intersection of each row and column must be a scalar value, and a table must not contain any repeating columns. |
| Second (2NF) | Every non-key column must depend on the entire primary key. |
| Third (3NF) | Every non-key column must depend only on the primary key. |

# Data Normalization

- The next four normal forms

| Normal form | Description |
|---|---|
| Boyce-Codd (BCNF) | A non-key column can't be dependent on another non-key column. |
| Fourth (4NF) | A table must not have more than one *multivalued dependency*, where the primary key has a one-to-many relationship to non-key columns. |
| Fifth (5NF) | The data structure is split into smaller and smaller tables until all redundancy has been eliminated. |
| Domain-key (DKNF) or Sixth (6NF) | Every constraint on the relationship is dependent only on key constraints and domain constraints, where a *domain* is the set of allowable values for a column. |

# Unnormalized Invoice Data

- Design with a column that contains multiple values

| vendor_name | invoice_number | item_description |
|---|---|---|
| Cahners Publishing | 112897 | VB ad, SQL ad, Library directory |
| Zylka Design | 97/522 | Catalogs, SQL Flyer |
| Zylka Design | 97/533B | Card revision |

- Design with multiple, repeated columns

| vendor_name | invoice_number | item_description_1 | item_description_2 | item_description_3 |
|---|---|---|---|---|
| Cahners Publishing | 112897 | VB ad | SQL ad | Library directory |
| Zylka Design | 97/552 | Catalogs | SQL flyer | NULL |
| Zylka Design | 97/553B | Card revision | NULL | NULL |

# Invoice Data in First Normal Form

| invoice_id | vendor_name | invoice_number | invoice_sequence | item_description |
|---|---|---|---|---|
| 1 | Cahners Publishing | 112897 | 1 | VB ad |
| 1 | Cahners Publishing | 112897 | 2 | SQL ad |
| 1 | Cahners Publishing | 112897 | 3 | Library directory |
| 2 | Zylka Design | 97/522 | 1 | Catalogs |
| 2 | Zylka Design | 97/522 | 2 | SQL flyer |
| 3 | Zylka Design | 97/533B | 1 | Card revision |

With primary key added

| | |
|---|---|
| First (1NF) | The value stored at the intersection of each row and column must be a scalar value, and a table must not contain any repeating columns. |

# Invoice Data in Second Normal Form

| invoice_id | vendor_name | invoice_number | invoice_sequence | item_description |
|---|---|---|---|---|
| 1 | Cahners Publishing | 112897 | 1 | VB ad |
| 1 | Cahners Publishing | 112897 | 2 | SQL ad |
| 1 | Cahners Publishing | 112897 | 3 | Library directory |
| 2 | Zylka Design | 97/522 | 1 | Catalogs |
| 2 | Zylka Design | 97/522 | 2 | SQL flyer |
| 3 | Zylka Design | 97/533B | 1 | Card revision |

First (1NF):

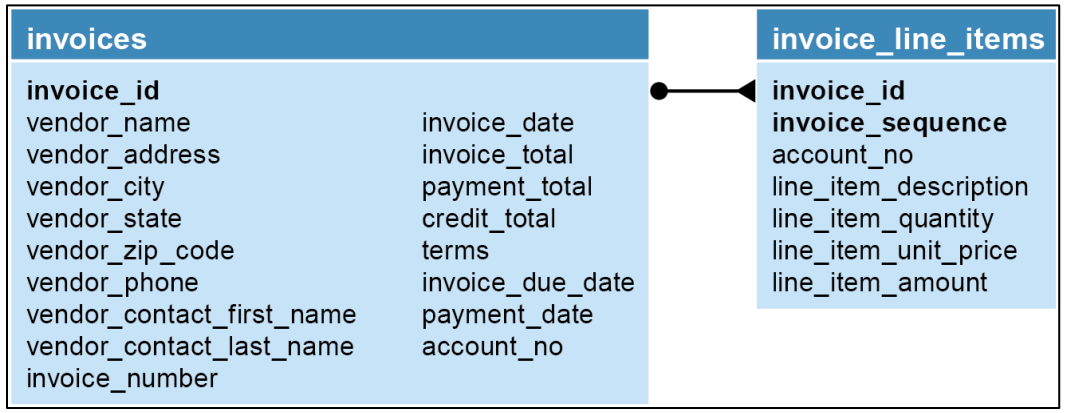
| invoice_number | vendor_name | invoice_id |
|---|---|---|
| 112897 | Cahners Publishing | 1 |
| 97/522 | Zylka Design | 2 |
| 97/533B | Zylka Design | 3 |

| invoice_id | invoice_sequence | item_description |
|---|---|---|
| 1 | 1 | VB ad |
| 1 | 2 | SQL ad |
| 1 | 3 | Library directory |
| 2 | 1 | Catalogs |
| 2 | 2 | SQL flyer |
| 3 | 1 | Card revision |

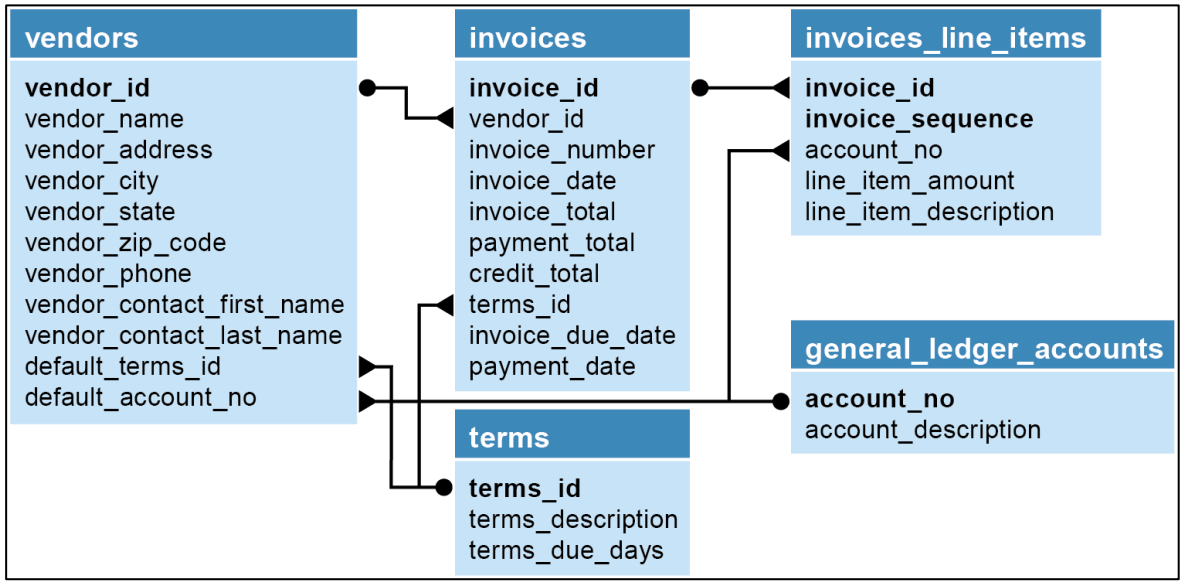
Second (2NF):

Every non-key column must depend on the entire primary key

# AP System



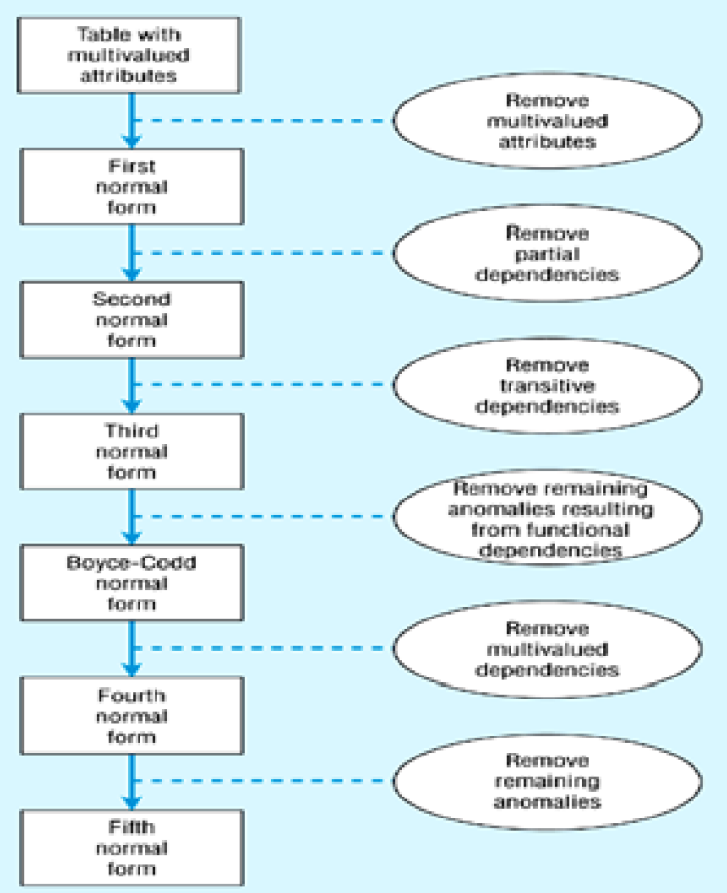

Second Normal Form



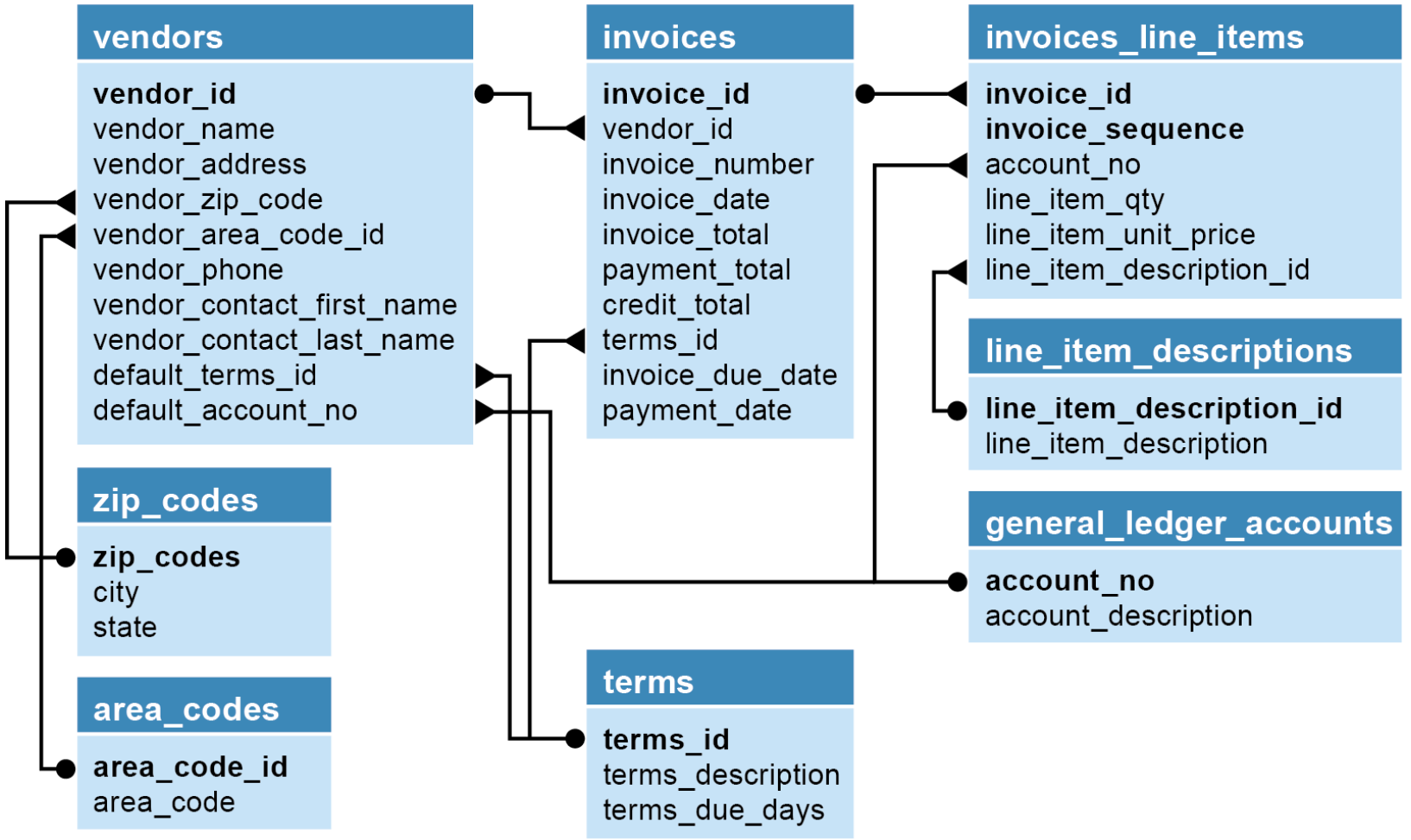

Third Normal Form

| | |
|---|---|
| Third (3NF) | Every non-key column must depend only on the primary key. |

# A Nice Data Normalization Summary



# AP System in Fifth Normal Form



# Denormalizing Data

- Benefits of well-normalized data
  - Minimize duplicate data
  - Efficiency in data maintenance/modification, minimize opportunities for data inconsistency
  - Simplify queries
- Possible situations to “denormalize”
  - A column from a joined table is used repeatedly in search criteria
  - A table is updated infrequently
  - Derived columns are used frequently in search conditions
  - More…use your judgement!
- “Normalize till it hurts, denormalize till it works”

# Billboard Music Data

Raw data file: Billboard Top 100
URL: https://raw.githubusercontent.com/hadley/tidy-data/master/data/billboard.csv
Raw data description: Weekly rank for songs on Billboard's Top 100 list.

| Field | Description |
|---|---|
| **Year** | Year |
| **Artist** | Artist name (last, first) |
| **Track** | Track name |
| **Time** | Track time |
| **Genre** | Track genre |
| **Date Entered** | Date first appearing on Top 100 |
| **Date Peaked** | Date of peak on Top 100 |
| **1st week … 76th week** | Rank on Top 100 in week 1…76 |

# Billboard Music Data

| | year | artist | track | time | genre | date.entered | date.peaked | wk1 | wk2 | wk3 | wk4 | wk5 | wk6 | wk7 | wk8 | wk9 | wk10 | wk11 | … | wk75 | wk76 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 225 | 2000 | Next | Wifey | 4:03 | Rock | 2000-05-27 | 2000-09-09 | 85 | 61 | 46 | 40 | 36 | 31 | 25 | 22 | 25 | 24 | 21 | … | NA | NA |
| 226 | 2000 | Nine Days | Absolutely | 3:09 | Rock | 2000-05-06 | 2000-07-22 | 85 | 71 | 59 | 52 | 39 | 34 | 26 | 20 | 17 | 13 | 11 | … | NA | NA |
| 227 | 2000 | Nine Days | If I Am | 4:18 | Rock | 2000-12-02 | 2000-12-02 | 68 | 68 | 81 | 94 | 100 | NA | NA | NA | NA | NA | NA | … | NA | NA |
| 228 | 2000 | No Doubt | Simple Kind Of Life | 4:11 | Rock | 2000-07-01 | 2000-07-22 | 50 | 40 | 39 | 38 | 38 | 48 | 52 | 55 | 80 | 85 | 88 | … | NA | NA |
| 229 | 2000 | Nu Flavor | 3 Little Words | 3:54 | R&B | 2000-06-03 | 2000-06-17 | 97 | 97 | 89 | 89 | 94 | 90 | 99 | 97 | 98 | NA | NA | … | NA | NA |
| 230 | 2000 | Offspring, The | Original Prankster | 3:42 | Rock | 2000-11-25 | 2000-12-09 | 74 | 71 | 70 | 70 | 77 | 78 | 91 | 89 | 92 | NA | NA | … | NA | NA |
| 231 | 2000 | Paisley, Brad | Me Neither | 3:19 | Country | 2000-05-13 | 2000-05-20 | 87 | 85 | 90 | 92 | NA | NA | NA | NA | NA | NA | NA | … | NA | NA |
| 232 | 2000 | Paisley, Brad | We Danced | 3:45 | Country | 2000-10-14 | 2000-12-16 | 71 | 68 | 52 | 52 | 45 | 42 | 39 | 34 | 34 | 29 | 36 | … | NA | NA |
| 233 | 2000 | Papa Roach | Last Resort | 3:19 | Rock | 2000-07-29 | 2000-12-02 | 75 | 71 | 69 | 69 | 66 | 64 | 61 | 61 | 66 | 64 | 63 | … | NA | NA |
| 234 | 2000 | Pearl Jam | Nothing As It Seems | 5:20 | Rock | 2000-05-13 | 2000-05-13 | 49 | 70 | 84 | 89 | 93 | 91 | NA | NA | NA | NA | NA | … | NA | NA |
| 235 | 2000 | Pink | Most Girls | 4:06 | Rock | 2000-08-12 | 2000-11-25 | 85 | 70 | 52 | 36 | 27 | 21 | 15 | 13 | 12 | 8 | 5 | … | NA | NA |
| 236 | 2000 | Pink | There U Go | 3:23 | Rock | 2000-03-04 | 2000-04-08 | 25 | 15 | 12 | 11 | 11 | 7 | 7 | 12 | 14 | 15 | 15 | … | NA | NA |
| 237 | 2000 | Price, Kelly | As We Lay | 6:20 | Rock | 2000-07-15 | 2000-08-05 | 82 | 69 | 69 | 64 | 71 | 79 | 82 | 89 | NA | NA | NA | … | NA | NA |
| 238 | 2000 | Price, Kelly | Love Sets You Free | 3:46 | Rock | 2000-05-13 | 2000-05-20 | 92 | 91 | 98 | 100 | NA | NA | NA | NA | NA | NA | NA | … | NA | NA |
| 239 | 2000 | Price, Kelly | You Should've Told Me | 3:12 | Rock | 2000-09-23 | 2000-12-02 | 91 | 91 | 91 | 87 | 86 | 79 | 79 | 75 | 70 | 68 | 64 | … | NA | NA |
| 240 | 2000 | Profyle | Liar | 3:57 | R&B | 2000-09-16 | 2000-10-28 | 52 | 32 | 25 | 17 | 16 | 16 | 14 | 19 | 24 | 33 | 35 | … | NA | NA |
| 241 | 2000 | Puff Daddy | Best Friend | 5:33 | Rap | 2000-02-12 | 2000-02-19 | 65 | 59 | 62 | 79 | 99 | NA | NA | NA | NA | NA | NA | … | NA | NA |
| 242 | 2000 | Q-Tip | Breathe And Stop | 4:06 | Rock | 2000-01-22 | 2000-01-22 | 71 | 71 | 81 | 82 | 96 | NA | NA | NA | NA | NA | NA | … | NA | NA |
| 243 | 2000 | R.E.M. | The Great Beyond | 4:10 | Rock | 1999-12-25 | 2000-01-29 | 79 | 79 | 70 | 62 | 60 | 57 | 61 | 66 | 60 | 59 | 60 | … | NA | NA |
| 244 | 2000 | Rascal Flatts | Prayin' For Daylight | 3:36 | Country | 2000-05-06 | 2000-08-12 | 87 | 78 | 72 | 68 | 66 | 64 | 58 | 58 | 56 | 54 | 49 | … | NA | NA |
| 245 | 2000 | Raye, Collin | Couldn't Last A Moment | 3:40 | Country | 2000-03-18 | 2000-06-24 | 91 | 85 | 75 | 73 | 67 | 63 | 63 | 63 | 56 | 49 | 48 | … | NA | NA |
| 246 | 2000 | Red Hot Chili Peppers | Californication | 5:21 | Rock | 2000-07-29 | 2000-10-14 | 72 | 72 | 72 | 77 | 79 | 77 | 75 | 84 | 85 | 79 | 74 | … | NA | NA |
| 247 | 2000 | Red Hot Chili Peppers | Otherside | 4:13 | Rock | 2000-02-12 | 2000-05-27 | 80 | 72 | 65 | 52 | 51 | 49 | 40 | 37 | 32 | 29 | 29 | … | NA | NA |
| 248 | 2000 | Rimes, LeAnn | Big Deal | 3:03 | Country | 1999-10-16 | 2000-01-01 | 71 | 52 | 51 | 51 | 51 | 48 | 41 | 37 | 29 | 26 | 26 | … | NA | NA |
| 249 | 2000 | Rimes, LeAnn | Can't Fight The Moonlight | 3:33 | Country | 2000-09-09 | 2000-09-16 | 82 | 71 | 79 | 83 | 96 | 99 | 78 | 78 | 83 | 79 | 77 | … | NA | NA |
| 250 | 2000 | Rimes, LeAnn | I Need You | 3:42 | Country | 2000-05-27 | 2000-08-12 | 77 | 68 | 67 | 63 | 59 | 59 | 59 | 53 | 51 | 50 | 13 | … | NA | NA |

# Billboard Music Data

| | year | artist | track | time | genre | date.entered | date.peaked | week | rank |
|---|---|---|---|---|---|---|---|---|---|
| 3860 | 2000 | Nine Days | If I Am | 4:18 | Rock | 2000-12-02 | 2000-12-02 | 1 | 68 |
| 3861 | 2000 | Nine Days | If I Am | 4:18 | Rock | 2000-12-02 | 2000-12-02 | 2 | 68 |
| 3862 | 2000 | Nine Days | If I Am | 4:18 | Rock | 2000-12-02 | 2000-12-02 | 3 | 81 |
| 3863 | 2000 | Nine Days | If I Am | 4:18 | Rock | 2000-12-02 | 2000-12-02 | 4 | 94 |
| 3864 | 2000 | Nine Days | If I Am | 4:18 | Rock | 2000-12-02 | 2000-12-02 | 5 | 100 |
| 3865 | 2000 | No Doubt | Simple Kind Of Life | 4:11 | Rock | 2000-07-01 | 2000-07-22 | 1 | 50 |
| 3866 | 2000 | No Doubt | Simple Kind Of Life | 4:11 | Rock | 2000-07-01 | 2000-07-22 | 2 | 40 |
| 3867 | 2000 | No Doubt | Simple Kind Of Life | 4:11 | Rock | 2000-07-01 | 2000-07-22 | 3 | 39 |
| 3868 | 2000 | No Doubt | Simple Kind Of Life | 4:11 | Rock | 2000-07-01 | 2000-07-22 | 4 | 38 |
| 3869 | 2000 | No Doubt | Simple Kind Of Life | 4:11 | Rock | 2000-07-01 | 2000-07-22 | 5 | 38 |
| 3870 | 2000 | No Doubt | Simple Kind Of Life | 4:11 | Rock | 2000-07-01 | 2000-07-22 | 6 | 48 |
| 3871 | 2000 | No Doubt | Simple Kind Of Life | 4:11 | Rock | 2000-07-01 | 2000-07-22 | 7 | 52 |
| 3872 | 2000 | No Doubt | Simple Kind Of Life | 4:11 | Rock | 2000-07-01 | 2000-07-22 | 8 | 55 |
| 3873 | 2000 | No Doubt | Simple Kind Of Life | 4:11 | Rock | 2000-07-01 | 2000-07-22 | 9 | 80 |
| 3874 | 2000 | No Doubt | Simple Kind Of Life | 4:11 | Rock | 2000-07-01 | 2000-07-22 | 10 | 85 |
| 3875 | 2000 | No Doubt | Simple Kind Of Life | 4:11 | Rock | 2000-07-01 | 2000-07-22 | 11 | 88 |
| 3876 | 2000 | No Doubt | Simple Kind Of Life | 4:11 | Rock | 2000-07-01 | 2000-07-22 | 12 | 100 |
| 3877 | 2000 | No Doubt | Simple Kind Of Life | 4:11 | Rock | 2000-07-01 | 2000-07-22 | 13 | 98 |
| 3878 | 2000 | Nu Flavor | 3 Little Words | 3:54 | R&B | 2000-06-03 | 2000-06-17 | 1 | 97 |
| 3879 | 2000 | Nu Flavor | 3 Little Words | 3:54 | R&B | 2000-06-03 | 2000-06-17 | 2 | 97 |
| 3880 | 2000 | Nu Flavor | 3 Little Words | 3:54 | R&B | 2000-06-03 | 2000-06-17 | 3 | 89 |
| 3881 | 2000 | Nu Flavor | 3 Little Words | 3:54 | R&B | 2000-06-03 | 2000-06-17 | 4 | 89 |
| 3882 | 2000 | Nu Flavor | 3 Little Words | 3:54 | R&B | 2000-06-03 | 2000-06-17 | 5 | 94 |
| 3883 | 2000 | Nu Flavor | 3 Little Words | 3:54 | R&B | 2000-06-03 | 2000-06-17 | 6 | 90 |
| 3884 | 2000 | Nu Flavor | 3 Little Words | 3:54 | R&B | 2000-06-03 | 2000-06-17 | 7 | 99 |
| 3885 | 2000 | Nu Flavor | 3 Little Words | 3:54 | R&B | 2000-06-03 | 2000-06-17 | 8 | 97 |

Repeating columns (wk1, wk2, …wk76) transformed to rows with single “week” column

# Billboard Music Data

| songID | year | artist | track | time | genre | date.entered | date.peaked | week | rank |
|---|---|---|---|---|---|---|---|---|---|
| 227 | 2000 | Nine Days | If I Am | 4:18 | Rock | 2000-12-02 | 2000-12-02 | 1 | 68 |
| 227 | 2000 | Nine Days | If I Am | 4:18 | Rock | 2000-12-02 | 2000-12-02 | 2 | 68 |
| 227 | 2000 | Nine Days | If I Am | 4:18 | Rock | 2000-12-02 | 2000-12-02 | 3 | 81 |
| 227 | 2000 | Nine Days | If I Am | 4:18 | Rock | 2000-12-02 | 2000-12-02 | 4 | 94 |
| 227 | 2000 | Nine Days | If I Am | 4:18 | Rock | 2000-12-02 | 2000-12-02 | 5 | 100 |
| 228 | 2000 | No Doubt | Simple Kind Of Life | 4:11 | Rock | 2000-07-01 | 2000-07-22 | 1 | 50 |
| 228 | 2000 | No Doubt | Simple Kind Of Life | 4:11 | Rock | 2000-07-01 | 2000-07-22 | 2 | 40 |
| 228 | 2000 | No Doubt | Simple Kind Of Life | 4:11 | Rock | 2000-07-01 | 2000-07-22 | 3 | 39 |
| 228 | 2000 | No Doubt | Simple Kind Of Life | 4:11 | Rock | 2000-07-01 | 2000-07-22 | 4 | 38 |
| 228 | 2000 | No Doubt | Simple Kind Of Life | 4:11 | Rock | 2000-07-01 | 2000-07-22 | 5 | 38 |
| 228 | 2000 | No Doubt | Simple Kind Of Life | 4:11 | Rock | 2000-07-01 | 2000-07-22 | 6 | 48 |
| 228 | 2000 | No Doubt | Simple Kind Of Life | 4:11 | Rock | 2000-07-01 | 2000-07-22 | 7 | 52 |
| 228 | 2000 | No Doubt | Simple Kind Of Life | 4:11 | Rock | 2000-07-01 | 2000-07-22 | 8 | 55 |
| 228 | 2000 | No Doubt | Simple Kind Of Life | 4:11 | Rock | 2000-07-01 | 2000-07-22 | 9 | 80 |
| 228 | 2000 | No Doubt | Simple Kind Of Life | 4:11 | Rock | 2000-07-01 | 2000-07-22 | 10 | 85 |
| 228 | 2000 | No Doubt | Simple Kind Of Life | 4:11 | Rock | 2000-07-01 | 2000-07-22 | 11 | 88 |
| 228 | 2000 | No Doubt | Simple Kind Of Life | 4:11 | Rock | 2000-07-01 | 2000-07-22 | 12 | 100 |
| 228 | 2000 | No Doubt | Simple Kind Of Life | 4:11 | Rock | 2000-07-01 | 2000-07-22 | 13 | 98 |
| 229 | 2000 | Nu Flavor | 3 Little Words | 3:54 | R&B | 2000-06-03 | 2000-06-17 | 1 | 97 |
| 229 | 2000 | Nu Flavor | 3 Little Words | 3:54 | R&B | 2000-06-03 | 2000-06-17 | 2 | 97 |
| 229 | 2000 | Nu Flavor | 3 Little Words | 3:54 | R&B | 2000-06-03 | 2000-06-17 | 3 | 89 |
| 229 | 2000 | Nu Flavor | 3 Little Words | 3:54 | R&B | 2000-06-03 | 2000-06-17 | 4 | 89 |
| 229 | 2000 | Nu Flavor | 3 Little Words | 3:54 | R&B | 2000-06-03 | 2000-06-17 | 5 | 94 |
| 229 | 2000 | Nu Flavor | 3 Little Words | 3:54 | R&B | 2000-06-03 | 2000-06-17 | 6 | 90 |
| 229 | 2000 | Nu Flavor | 3 Little Words | 3:54 | R&B | 2000-06-03 | 2000-06-17 | 7 | 99 |
| 229 | 2000 | Nu Flavor | 3 Little Words | 3:54 | R&B | 2000-06-03 | 2000-06-17 | 8 | 97 |

“songID” identifier added, primary key = ???

# Billboard Music Data

| songID | artist | track | genre | time |
|---|---|---|---|---|
| 227 | Nine Days | If I Am | Rock | 4:18 |
| 228 | No Doubt | Simple Kind Of Life | Rock | 4:11 |
| 229 | Nu Flavor | 3 Little Words | R&B | 3:54 |
| 230 | Offspring, The | Original Prankster | Rock | 3:42 |
| 231 | Paisley, Brad | Me Neither | Country | 3:19 |
| 232 | Paisley, Brad | We Danced | Country | 3:45 |
| 233 | Papa Roach | Last Resort | Rock | 3:19 |

| songID | date.entered | date.peaked | week | rank |
|---|---|---|---|---|
| 227 | 2000-12-02 | 2000-12-02 | 1 | 68 |
| 227 | 2000-12-02 | 2000-12-02 | 2 | 68 |
| 227 | 2000-12-02 | 2000-12-02 | 3 | 81 |
| 227 | 2000-12-02 | 2000-12-02 | 4 | 94 |
| 227 | 2000-12-02 | 2000-12-02 | 5 | 100 |
| 228 | 2000-07-01 | 2000-07-22 | 1 | 50 |
| 228 | 2000-07-01 | 2000-07-22 | 2 | 40 |
| 228 | 2000-07-01 | 2000-07-22 | 3 | 39 |
| 228 | 2000-07-01 | 2000-07-22 | 4 | 38 |
| 228 | 2000-07-01 | 2000-07-22 | 5 | 38 |
| 228 | 2000-07-01 | 2000-07-22 | 6 | 48 |
| 228 | 2000-07-01 | 2000-07-22 | 7 | 52 |
| 228 | 2000-07-01 | 2000-07-22 | 8 | 55 |
| 228 | 2000-07-01 | 2000-07-22 | 9 | 80 |
| 228 | 2000-07-01 | 2000-07-22 | 10 | 85 |
| 228 | 2000-07-01 | 2000-07-22 | 11 | 88 |
| 228 | 2000-07-01 | 2000-07-22 | 12 | 100 |
| 228 | 2000-07-01 | 2000-07-22 | 13 | 98 |
| 229 | 2000-06-03 | 2000-06-17 | 1 | 97 |
| 229 | 2000-06-03 | 2000-06-17 | 2 | 97 |
| 229 | 2000-06-03 | 2000-06-17 | 3 | 89 |
| 229 | 2000-06-03 | 2000-06-17 | 4 | 89 |
| 229 | 2000-06-03 | 2000-06-17 | 5 | 94 |
| 229 | 2000-06-03 | 2000-06-17 | 6 | 90 |
| 229 | 2000-06-03 | 2000-06-17 | 7 | 99 |
| 229 | 2000-06-03 | 2000-06-17 | 8 | 97 |

Move columns that do not depend on entire primary key to separate table

Now we have a “Songs” table and a “Weeks” table

Primary keys = ???, foreign keys = ???

# Billboard Music Data

| songID | artist | track | genre | time |
|---|---|---|---|---|
| 227 | Nine Days | If I Am | Rock | 4:18 |
| 228 | No Doubt | Simple Kind Of Life | Rock | 4:11 |
| 229 | Nu Flavor | 3 Little Words | R&B | 3:54 |
| 230 | Offspring, The | Original Prankster | Rock | 3:42 |
| 231 | Paisley, Brad | Me Neither | Country | 3:19 |
| 232 | Paisley, Brad | We Danced | Country | 3:45 |
| 233 | Papa Roach | Last Resort | Rock | 3:19 |

Clean up date/week handling so we have more easily usable fields

| songID | date | week | rank |
|---|---|---|---|
| 227 | 2000-12-02 | 1 | 68 |
| 227 | 2000-12-09 | 2 | 68 |
| 227 | 2000-12-16 | 3 | 81 |
| 227 | 2000-12-23 | 4 | 94 |
| 227 | 2000-12-30 | 5 | 100 |
| 228 | 2000-07-01 | 1 | 50 |
| 228 | 2000-07-08 | 2 | 40 |
| 228 | 2000-07-15 | 3 | 39 |
| 228 | 2000-07-22 | 4 | 38 |
| 228 | 2000-07-29 | 5 | 38 |
| 228 | 2000-08-05 | 6 | 48 |
| 228 | 2000-08-12 | 7 | 52 |
| 228 | 2000-08-19 | 8 | 55 |
| 228 | 2000-08-26 | 9 | 80 |
| 228 | 2000-09-02 | 10 | 85 |
| 228 | 2000-09-09 | 11 | 88 |
| 228 | 2000-09-16 | 12 | 100 |
| 228 | 2000-09-23 | 13 | 98 |
| 229 | 2000-06-03 | 1 | 97 |
| 229 | 2000-06-10 | 2 | 97 |
| 229 | 2000-06-17 | 3 | 89 |
| 229 | 2000-06-24 | 4 | 89 |
| 229 | 2000-07-01 | 5 | 94 |
| 229 | 2000-07-08 | 6 | 90 |
| 229 | 2000-07-15 | 7 | 99 |
| 229 | 2000-07-22 | 8 | 97 |

# Some Queries of Billboard Music Data

```
1  use music;
2
3  -- recreate TOp 100 for given date
4  SELECT date, top100rank, weeks.songID, track, artist
5  FROM weeks
6      JOIN songs ON weeks.songID = songs.songID
7  WHERE date = "2000-08-12"
8  ORDER BY top100rank;
9
10 -- find weeks of peak for given song
11 SELECT date, weeks.songID, track, artist, top100rank AS "peak_rank"
12 FROM weeks
13     JOIN songs ON weeks.songID = songs.songID
14 WHERE weeks.songID = 228 AND
15         top100rank = (SELECT MIN(top100rank) AS "peak_rank" FROM weeks WHERE weeks.songID = 228)
16 ORDER BY date;
17
18 -- tracks on the list for more than 12 weeks
19 SELECT date, weeks.songID, track, artist, MAX(week) AS "time_on_top100"
20 FROM weeks
21     JOIN songs ON weeks.songID = songs.songID
22 GROUP BY weeks.songID
23 HAVING time_on_top100 > 12
24 ORDER BY time_on_top100 DESC;
25
26
```

| date | top100rank | songID | track | artist |
|---|---|---|---|---|
| 2000-08-12 | 1 | 265 | Incomplete | Sisqo |
| 2000-08-12 | 2 | 198 | Bent | matchbox twenty |
| 2000-08-12 | 3 | 221 | It's Gonna Be Me | N'Sync |
| 2000-08-12 | 4 | 73 | Jumpin' Jumpin' | Destiny's Child |
| 2000-08-12 | 5 | 9 | Try Again | Aaliyah |
| 2000-08-12 | 6 | 150 | I Wanna Know | Joe |
| 2000-08-12 | 7 | 302 | Everything You Want | Vertical Horizon |
| 2000-08-12 | 8 | 226 | Absolutely | Nine Days |
| 2000-08-12 | 9 | 63 | Higher | Creed |
| 2000-08-12 | 10 | 143 | Doesn't Really Matter | Janet |
| 2000-08-12 | 11 | 250 | I Need You | Rimes, LeAnn |
| 2000-08-12 | 12 | 252 | No More | Ruff Endz |
| 2000-08-12 | 13 | 42 | He Wasn't Man Enou... | Braxton, Toni |
| 2000-08-12 | 14 | 142 | Let's Get Married | Jagged Edge |
| 2000-08-12 | 15 | 29 | Back Here | BBMak |
| 2000-08-12 | 16 | 236 | There U Go | Pink |
| 2000-08-12 | 17 | 224 | (Hot S**t) Country ... | Nelly |
| 2000-08-12 | 18 | 3 | Kryptonite | 3 Doors Down |

| date | songID | track | artist | peak_rank |
|---|---|---|---|---|
| 2000-07-22 | 228 | Simple Kind Of Life | No Doubt | 38 |
| 2000-07-29 | 228 | Simple Kind Of Life | No Doubt | 38 |

# Raw Data -> Data Ready for Analysis

## What you should deliver to the statistician

To facilitate the most efficient and timely analysis this is the information you should pass to a statistician:

1. The raw data.
2. A tidy data set
3. A code book describing each variable and its values in the tidy data set.
4. An explicit and exact recipe you used to go from 1 -> 2,3

https://github.com/jtleek/datasharing

### The tidy data set

The general principles of tidy data are laid out by Hadley Wickham in this paper and this video. While both the paper and the video describe tidy data using R, the principles are more generally applicable:

1. Each variable you measure should be in one column
2. Each different observation of that variable should be in a different row
3. There should be one table for each "kind" of variable
4. If you have multiple tables, they should include a column in the table that allows them to be joined or merged

https://www.jstatsoft.org/article/view/v059i10

# “Tidy” Data

## 3. Tidying messy datasets

Real datasets can, and often do, violate the three precepts of tidy data in almost every way imaginable. While occasionally you do get a dataset that you can start analyzing immediately, this is the exception, not the rule. This section describes the five most common problems with messy datasets, along with their remedies:

- 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.

https://www.jstatsoft.org/article/view/v059i10

# Values stored in column names

| religion | <$10k | $10–20k | $20–30k | $30–40k | $40–50k | $50–75k |
|---|---|---|---|---|---|---|
| Agnostic | 27 | 34 | 60 | 81 | 76 | 137 |
| Atheist | 12 | 27 | 37 | 52 | 35 | 70 |
| Buddhist | 27 | 21 | 30 | 34 | 33 | 58 |
| Catholic | 418 | 617 | 732 | 670 | 638 | 1116 |
| Don't know/refused | 15 | 14 | 15 | 11 | 10 | 35 |
| Evangelical Prot | 575 | 869 | 1064 | 982 | 881 | 1486 |
| Hindu | 1 | 9 | 7 | 9 | 11 | 34 |
| Historically Black Prot | 228 | 244 | 236 | 238 | 197 | 223 |
| Jehovah's Witness | 20 | 27 | 24 | 24 | 21 | 30 |
| Jewish | 19 | 19 | 25 | 25 | 30 | 95 |

Table 4: The first ten rows of data on income and religion from the Pew Forum. Three columns, $75–100k, $100–150k and >150k, have been omitted.

| religion | income | freq |
|---|---|---|
| Agnostic | <$10k | 27 |
| Agnostic | $10–20k | 34 |
| Agnostic | $20–30k | 60 |
| Agnostic | $30–40k | 81 |
| Agnostic | $40–50k | 76 |
| Agnostic | $50–75k | 137 |
| Agnostic | $75–100k | 122 |
| Agnostic | $100–150k | 109 |
| Agnostic | >150k | 84 |
| Agnostic | Don't know/refused | 96 |

Table 6: The first ten rows of the tidied Pew survey dataset on income and religion. The `column` has been renamed to `income`, and `value` to `freq`.

# Multiple variables stored in one column

| country | year | m014 | m1524 | m2534 | m3544 | m4554 | m5564 | m65 | mu | f014 |
|---|---|---|---|---|---|---|---|---|---|---|
| AD | 2000 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | — | — |
| AE | 2000 | 2 | 4 | 4 | 6 | 5 | 12 | 10 | — | 3 |
| AF | 2000 | 52 | 228 | 183 | 149 | 129 | 94 | 80 | — | 93 |
| AG | 2000 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | — | 1 |
| AL | 2000 | 2 | 19 | 21 | 14 | 24 | 19 | 16 | — | 3 |
| AM | 2000 | 2 | 152 | 130 | 131 | 63 | 26 | 21 | — | 1 |
| AN | 2000 | 0 | 0 | 1 | 2 | 0 | 0 | 0 | — | 0 |
| AO | 2000 | 186 | 999 | 1003 | 912 | 482 | 312 | 194 | — | 247 |
| AR | 2000 | 97 | 278 | 594 | 402 | 419 | 368 | 330 | — | 121 |
| AS | 2000 | — | — | — | — | 1 | 1 | — | — | — |

Table 9: Original TB dataset. Corresponding to each 'm' column for males, there is also an 'f' column for females, `f1524`, `f2534` and so on. These are not shown to conserve space. Note the mixture of 0s and missing values (—). This is due to the data collection process and the distinction is important for this dataset.

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

(a) Molten data

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

(b) Tidy data

Table 10: Tidying the TB dataset requires first melting, and then splitting the `column` column into two variables: `sex` and `age`.

Column headers in this format are often separated by some character (`.`, `-`, `_`, `:`). While the string can be broken into pieces using that character as a divider, in other cases, such as for this dataset, more careful string processing is required. For example, the variable names can be matched to a lookup table that converts single compound value into multiple component values.

# Variables in both rows and columns

| id | year | month | element | d1 | d2 | d3 | d4 | d5 | d6 | d7 | d8 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| MX17004 | 2010 | 1 | tmax | — | — | — | — | — | — | — | — |
| MX17004 | 2010 | 1 | tmin | — | — | — | — | — | — | — | — |
| MX17004 | 2010 | 2 | tmax | — | 27.3 | 24.1 | — | — | — | — | — |
| MX17004 | 2010 | 2 | tmin | — | 14.4 | 14.4 | — | — | — | — | — |
| MX17004 | 2010 | 3 | tmax | — | — | — | — | 32.1 | — | — | — |
| MX17004 | 2010 | 3 | tmin | — | — | — | — | 14.2 | — | — | — |
| MX17004 | 2010 | 4 | tmax | — | — | — | — | — | — | — | — |
| MX17004 | 2010 | 4 | tmin | — | — | — | — | — | — | — | — |
| MX17004 | 2010 | 5 | tmax | — | — | — | — | — | — | — | — |
| MX17004 | 2010 | 5 | tmin | — | — | — | — | — | — | — | — |

Table 11: Original weather dataset. There is a column for each possible day in the month. Columns d9 to d31 have been omitted to conserve space.

| id | date | element | value |
|---|---|---|---|
| MX17004 | 2010-01-30 | tmax | 27.8 |
| MX17004 | 2010-01-30 | tmin | 14.5 |
| MX17004 | 2010-02-02 | tmax | 27.3 |
| MX17004 | 2010-02-02 | tmin | 14.4 |
| MX17004 | 2010-02-03 | tmax | 24.1 |
| MX17004 | 2010-02-03 | tmin | 14.4 |
| MX17004 | 2010-02-11 | tmax | 29.7 |
| MX17004 | 2010-02-11 | tmin | 13.4 |
| MX17004 | 2010-02-23 | tmax | 29.9 |
| MX17004 | 2010-02-23 | tmin | 10.7 |

(a) Molten data

| id | date | tmax | tmin |
|---|---|---|---|
| MX17004 | 2010-01-30 | 27.8 | 14.5 |
| MX17004 | 2010-02-02 | 27.3 | 14.4 |
| MX17004 | 2010-02-03 | 24.1 | 14.4 |
| MX17004 | 2010-02-11 | 29.7 | 13.4 |
| MX17004 | 2010-02-23 | 29.9 | 10.7 |
| MX17004 | 2010-03-05 | 32.1 | 14.2 |
| MX17004 | 2010-03-10 | 34.5 | 16.8 |
| MX17004 | 2010-03-16 | 31.1 | 17.6 |
| MX17004 | 2010-04-27 | 36.3 | 16.7 |
| MX17004 | 2010-05-27 | 33.2 | 18.2 |

(b) Tidy data

Table 12: (a) Molten weather dataset. This is almost tidy, but instead of values, the `element` column contains names of variables. Missing values are dropped to conserve space. (b) Tidy weather dataset. Each row represents the meteorological measurements for a single day. There are two measured variables, minimum (`tmin`) and maximum (`tmax`) temperature; all other variables are fixed.

# Database Design Wrap-Up

1) Identify the data elements
    1) Look at real world model/documents
    2) Draft ER diagram from specification
    3) Etc…
2) Subdivide/group elements into entities with attributes
3) Identify relationships between entities
4) Normalize
5) Convert to relational schema (tables, columns, keys, indexes)