# TIDY DATA

MPA 630: Data Science for Public Management

September 20, 2018

**Fill out your reading report on Learning Suite**

# PLAN FOR TODAY

**Tidy data**

**Verbs**

**Live example**

---

**Birds and planes again**

# TIDY DATA

# WHAT IS TIDY DATA?

~~Clean perfect data~~

**Each variable is a column**

**Each observation is a row**

# COLUMN HEADERS ARE VALUES

| Country | Beer servings | Wine servings | Spirit servings |
|---|---|---|---|
| Canada | 240 | 122 | 100 |
| South Korea | 140 | 16 | 9 |
| USA | 249 | 128 | 84 |

# TIDIED

| Country | Type | Servings |
|---|---|---|
| Canada | Beer | 240 |
| South Korea | Beer | 140 |
| USA | Beer | 249 |
| Canada | Wine | 122 |
| South Korea | Wine | 16 |
| USA | Wine | 128 |
| Canada | Spirits | 100 |
| South Korea | Spirits | 9 |
| USA | Spirits | 84 |

# INFORMATION NOT IN TABLE

| Country | Beer servings | Wine servings | Spirit servings |
|---|---|---|---|
| Canada | **240** | 122 | 100 |
| South Korea | 140 | *16* | *9* |
| USA | **249** | 128 | 84 |

**Key**

North America

Asia

**Surprisingly high**

*Surprisingly low*

# TIDIED

| Country | Type | Servings | Continent | Surprise |
|---|---|---|---|---|
| Canada | Beer | 240 | North America | High |
| South Korea | Beer | 140 | Asia | NA |
| USA | Beer | 249 | North America | High |
| Canada | Wine | 122 | North America | NA |
| South Korea | Wine | 16 | Asia | Low |
| USA | Wine | 128 | North America | NA |
| Canada | Spirits | 100 | North America | NA |
| South Korea | Spirits | 9 | Asia | Low |
| USA | Spirits | 84 | North America | NA |

# WIDE VS LONG

wide

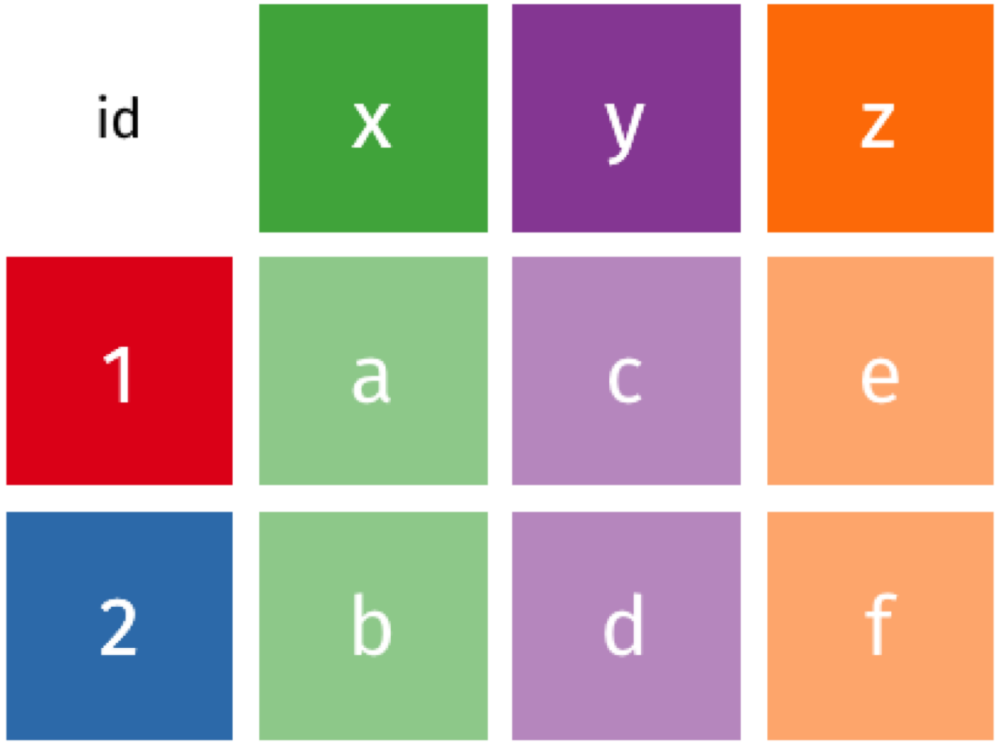
| id | x | y | z |
|---|---|---|---|
| 1 | a | c | e |
| 2 | b | d | f |

long

| id | key | val |
|---|---|---|
| 1 | x | a |
| 2 | x | b |
| 1 | y | c |
| 2 | y | d |
| 1 | z | e |
| 2 | z | f |

# MOVING FROM WIDE TO LONG

wide



# TIDY ME

| 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 |

# TIDIED

| 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 |

# TIDY ME TOO

| year | artist | track | time | date.entered | wk1 | wk2 | wk3 |
|---|---|---|---|---|---|---|---|
| 2000 | 2 Pac | Baby Don't Cry | 4:22 | 2000-02-26 | 87 | 82 | 72 |
| 2000 | 2Ge+her | The Hardest Part Of ... | 3:15 | 2000-09-02 | 91 | 87 | 92 |
| 2000 | 3 Doors Down | Kryptonite | 3:53 | 2000-04-08 | 81 | 70 | 68 |
| 2000 | 98^0 | Give Me Just One Nig... | 3:24 | 2000-08-19 | 51 | 39 | 34 |
| 2000 | A*Teens | Dancing Queen | 3:44 | 2000-07-08 | 97 | 97 | 96 |
| 2000 | Aaliyah | I Don't Wanna | 4:15 | 2000-01-29 | 84 | 62 | 51 |
| 2000 | Aaliyah | Try Again | 4:03 | 2000-03-18 | 59 | 53 | 38 |
| 2000 | Adams, Yolanda | Open My Heart | 5:30 | 2000-08-26 | 76 | 76 | 74 |

Table 7: The first eight Billboard top hits for 2000. Other columns not shown are `wk4`, `wk5`, ..., `wk75`.

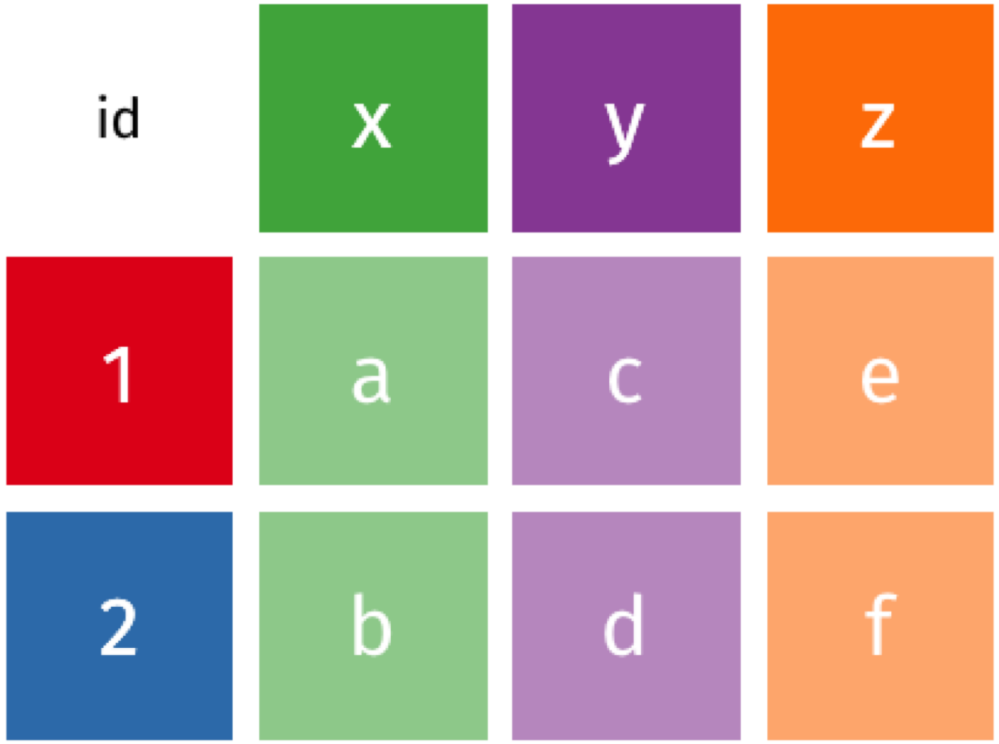
# TIDIED

| year | artist | time | track | date | week | rank |
|---|---|---|---|---|---|---|
| 2000 | 2 Pac | 4:22 | Baby Don't Cry | 2000-02-26 | 1 | 87 |
| 2000 | 2 Pac | 4:22 | Baby Don't Cry | 2000-03-04 | 2 | 82 |
| 2000 | 2 Pac | 4:22 | Baby Don't Cry | 2000-03-11 | 3 | 72 |
| 2000 | 2 Pac | 4:22 | Baby Don't Cry | 2000-03-18 | 4 | 77 |
| 2000 | 2 Pac | 4:22 | Baby Don't Cry | 2000-03-25 | 5 | 87 |
| 2000 | 2 Pac | 4:22 | Baby Don't Cry | 2000-04-01 | 6 | 94 |
| 2000 | 2 Pac | 4:22 | Baby Don't Cry | 2000-04-08 | 7 | 99 |
| 2000 | 2Ge+her | 3:15 | The Hardest Part Of ... | 2000-09-02 | 1 | 91 |
| 2000 | 2Ge+her | 3:15 | The Hardest Part Of ... | 2000-09-09 | 2 | 87 |
| 2000 | 2Ge+her | 3:15 | The Hardest Part Of ... | 2000-09-16 | 3 | 92 |
| 2000 | 3 Doors Down | 3:53 | Kryptonite | 2000-04-08 | 1 | 81 |
| 2000 | 3 Doors Down | 3:53 | Kryptonite | 2000-04-15 | 2 | 70 |
| 2000 | 3 Doors Down | 3:53 | Kryptonite | 2000-04-22 | 3 | 68 |
| 2000 | 3 Doors Down | 3:53 | Kryptonite | 2000-04-29 | 4 | 67 |
| 2000 | 3 Doors Down | 3:53 | Kryptonite | 2000-05-06 | 5 | 66 |

# VERBS

# GATHER AND SPREAD

wide



# MOST COMMON VERBS

| | |
|---|---|
| `filter()` | Choose rows based on conditions |
| `select()` | Choose (and rename) columns |
| `mutate()` | Add column (or change existing column) |
| `group_by()` | Make subgroups based on a column |
| `summarize()` | Calculate summary statistics for groups |

# FILTER

```
gapminder %>%
  filter(year == 1967)
```

| country <fctr> | continent <fctr> | year <int> | lifeExp <dbl> | pop <int> | gdpPercap <dbl> |
|---|---|---|---|---|---|
| Afghanistan | Asia | 1967 | 34.02000 | 11537966 | 836.1971 |
| Albania | Europe | 1967 | 66.22000 | 1984060 | 2760.1969 |
| Algeria | Africa | 1967 | 51.40700 | 12760499 | 3246.9918 |
| Angola | Africa | 1967 | 35.98500 | 5247469 | 5522.7764 |
| Argentina | Americas | 1967 | 65.63400 | 22934225 | 8052.9530 |
| Australia | Oceania | 1967 | 71.10000 | 11872264 | 14526.1246 |
| Austria | Europe | 1967 | 70.14000 | 7376998 | 12834.6024 |
| Bahrain | Asia | 1967 | 59.92300 | 202182 | 14804.6727 |
| Bangladesh | Asia | 1967 | 43.45300 | 62821884 | 721.1861 |
| Belgium | Europe | 1967 | 70.94000 | 9556500 | 13149.0412 |

1-10 of 142 rows

# FILTER

```
gapminder %>%
  filter(lifeExp < 40)
```

| country <fctr> | continent <fctr> | year <int> | lifeExp <dbl> | pop <int> | gdpPercap <dbl> |
|---|---|---|---|---|---|
| Afghanistan | Asia | 1952 | 28.801 | 8425333 | 779.4453 |
| Afghanistan | Asia | 1957 | 30.332 | 9240934 | 820.8530 |
| Afghanistan | Asia | 1962 | 31.997 | 10267083 | 853.1007 |
| Afghanistan | Asia | 1967 | 34.020 | 11537966 | 836.1971 |
| Afghanistan | Asia | 1972 | 36.088 | 13079460 | 739.9811 |
| Afghanistan | Asia | 1977 | 38.438 | 14880372 | 786.1134 |
| Afghanistan | Asia | 1982 | 39.854 | 12881816 | 978.0114 |
| Angola | Africa | 1952 | 30.015 | 4232095 | 3520.6103 |
| Angola | Africa | 1957 | 31.999 | 4561361 | 3827.9405 |
| Angola | Africa | 1962 | 34.000 | 4826015 | 4269.2767 |

1-10 of 124 rows

# FILTER

```
gapminder %>%
  filter(continent == "Asia", lifeExp < 40)
```

| country <fctr> | continent <fctr> | year <int> | lifeExp <dbl> | pop <int> | gdpPercap <dbl> |
|---|---|---|---|---|---|
| Afghanistan | Asia | 1952 | 28.801 | 8425333 | 779.4453 |
| Afghanistan | Asia | 1957 | 30.332 | 9240934 | 820.8530 |
| Afghanistan | Asia | 1962 | 31.997 | 10267083 | 853.1007 |
| Afghanistan | Asia | 1967 | 34.020 | 11537966 | 836.1971 |
| Afghanistan | Asia | 1972 | 36.088 | 13079460 | 739.9811 |
| Afghanistan | Asia | 1977 | 38.438 | 14880372 | 786.1134 |
| Afghanistan | Asia | 1982 | 39.854 | 12881816 | 978.0114 |
| Bangladesh | Asia | 1952 | 37.484 | 46886859 | 684.2442 |
| Bangladesh | Asia | 1957 | 39.348 | 51365468 | 661.6375 |
| Cambodia | Asia | 1952 | 39.417 | 4693836 | 368.4693 |

1-10 of 25 rows

Previous 1 2 3 Next

# SELECT

```
gapminder %>%
  select(country, year, pop)
```

| country <fctr> | year <int> | pop <int> |
|---|---|---|
| Afghanistan | 1952 | 8425333 |
| Afghanistan | 1957 | 9240934 |
| Afghanistan | 1962 | 10267083 |
| Afghanistan | 1967 | 11537966 |
| Afghanistan | 1972 | 13079460 |
| Afghanistan | 1977 | 14880372 |
| Afghanistan | 1982 | 12881816 |
| Afghanistan | 1987 | 13867957 |
| Afghanistan | 1992 | 16317921 |
| Afghanistan | 1997 | 22227415 |

1-10 of 1,704 rows

# MUTATE

```
gapminder %>%
  mutate(something_new = 5)
```

| country <fctr> | continent <fctr> | year <int> | lifeExp <dbl> | pop <int> | gdpPercap <dbl> | something_new <dbl> |
|---|---|---|---|---|---|---|
| Afghanistan | Asia | 1952 | 28.80100 | 8425333 | 779.4453 | 5 |
| Afghanistan | Asia | 1957 | 30.33200 | 9240934 | 820.8530 | 5 |
| Afghanistan | Asia | 1962 | 31.99700 | 10267083 | 853.1007 | 5 |
| Afghanistan | Asia | 1967 | 34.02000 | 11537966 | 836.1971 | 5 |
| Afghanistan | Asia | 1972 | 36.08800 | 13079460 | 739.9811 | 5 |
| Afghanistan | Asia | 1977 | 38.43800 | 14880372 | 786.1134 | 5 |
| Afghanistan | Asia | 1982 | 39.85400 | 12881816 | 978.0114 | 5 |
| Afghanistan | Asia | 1987 | 40.82200 | 13867957 | 852.3959 | 5 |
| Afghanistan | Asia | 1992 | 41.67400 | 16317921 | 649.3414 | 5 |
| Afghanistan | Asia | 1997 | 41.76300 | 22227415 | 635.3414 | 5 |

1-10 of 1,704 rows

# MUTATE

```
gapminder %>%
  mutate(pop_million = pop / 1000000)
```

| country <fctr> | continent <fctr> | year <int> | lifeExp <dbl> | pop <int> | gdpPercap <dbl> | pop_million <dbl> |
|---|---|---|---|---|---|---|
| Afghanistan | Asia | 1952 | 28.80100 | 8425333 | 779.4453 | 8.425333 |
| Afghanistan | Asia | 1957 | 30.33200 | 9240934 | 820.8530 | 9.240934 |
| Afghanistan | Asia | 1962 | 31.99700 | 10267083 | 853.1007 | 10.267083 |
| Afghanistan | Asia | 1967 | 34.02000 | 11537966 | 836.1971 | 11.537966 |
| Afghanistan | Asia | 1972 | 36.08800 | 13079460 | 739.9811 | 13.079460 |
| Afghanistan | Asia | 1977 | 38.43800 | 14880372 | 786.1134 | 14.880372 |
| Afghanistan | Asia | 1982 | 39.85400 | 12881816 | 978.0114 | 12.881816 |
| Afghanistan | Asia | 1987 | 40.82200 | 13867957 | 852.3959 | 13.867957 |
| Afghanistan | Asia | 1992 | 41.67400 | 16317921 | 649.3414 | 16.317921 |
| Afghanistan | Asia | 1997 | 41.76300 | 22227415 | 635.3414 | 22.227415 |

1-10 of 1,704 rows

# MUTATE

```
gapminder %>%
  mutate(lifeExp_binary = ifelse(lifeExp < 40,
                                 "Very low", "Not very low"))
```

| country <fctr> | continent <fctr> | year <int> | lifeExp <dbl> | pop <int> | gdpPercap <dbl> | lifeExp_binary <chr> |
|---|---|---|---|---|---|---|
| Afghanistan | Asia | 1952 | 28.80100 | 8425333 | 779.4453 | Very low |
| Afghanistan | Asia | 1957 | 30.33200 | 9240934 | 820.8530 | Very low |
| Afghanistan | Asia | 1962 | 31.99700 | 10267083 | 853.1007 | Very low |
| Afghanistan | Asia | 1967 | 34.02000 | 11537966 | 836.1971 | Very low |
| Afghanistan | Asia | 1972 | 36.08800 | 13079460 | 739.9811 | Very low |
| Afghanistan | Asia | 1977 | 38.43800 | 14880372 | 786.1134 | Very low |
| Afghanistan | Asia | 1982 | 39.85400 | 12881816 | 978.0114 | Very low |
| Afghanistan | Asia | 1987 | 40.82200 | 13867957 | 852.3959 | Not very low |
| Afghanistan | Asia | 1992 | 41.67400 | 16317921 | 649.3414 | Not very low |
| Afghanistan | Asia | 1997 | 41.76300 | 22227415 | 635.3414 | Not very low |

1-10 of 1,704 rows

# GROUP_BY + SUMMARIZE

```
gapminder %>%
  group_by(continent) %>%
  summarize(avg_lifeexp = mean(lifeExp),
            median_lifeexmp = median(lifeExp),
            num_countries = n())
```

| continent <fctr> | avg_lifeexp <dbl> | median_lifeexmp <dbl> | num_countries <int> |
|---|---|---|---|
| Africa | 48.86533 | 47.7920 | 624 |
| Americas | 64.65874 | 67.0480 | 300 |
| Asia | 60.06490 | 61.7915 | 396 |
| Europe | 71.90369 | 72.2410 | 360 |
| Oceania | 74.32621 | 73.6650 | 24 |

5 rows

# GROUP_BY + SUMMARIZE

```
gapminder %>%
  group_by(continent, year) %>%
  summarize(avg_lifeexp = mean(lifeExp),
            median_lifeexmp = median(lifeExp),
            num_countries = n())
```

| continent <fctr> | year <int> | avg_lifeexp <dbl> | median_lifeexmp <dbl> | num_countries <int> |
|---|---|---|---|---|
| Africa | 1952 | 39.13550 | 38.8330 | 52 |
| Africa | 1957 | 41.26635 | 40.5925 | 52 |
| Africa | 1962 | 43.31944 | 42.6305 | 52 |
| Africa | 1967 | 45.33454 | 44.6985 | 52 |
| Africa | 1972 | 47.45094 | 47.0315 | 52 |
| Africa | 1977 | 49.58042 | 49.2725 | 52 |
| Africa | 1982 | 51.59287 | 50.7560 | 52 |
| Africa | 1987 | 53.34479 | 51.6395 | 52 |
| Africa | 1992 | 53.62958 | 52.4290 | 52 |
| Africa | 1997 | 53.59827 | 52.7590 | 52 |

1-10 of 60 rows

Previous 1 2 3 4 5 6 Next

# OTHER HELPFUL VERBS

| | |
|---|---|
| `arrange()` | **Sort a data frame by a column** |
| `rename()` | **Rename columns** |
| `count()` | `group_by() %>% summarize(n = n())` |
| `gather()` | **Make a data frame long** |
| `spread()` | **Make a data frame wide** |

# LIVE EXAMPLE

# BREAK

# BIRDS AND PLANES AGAIN