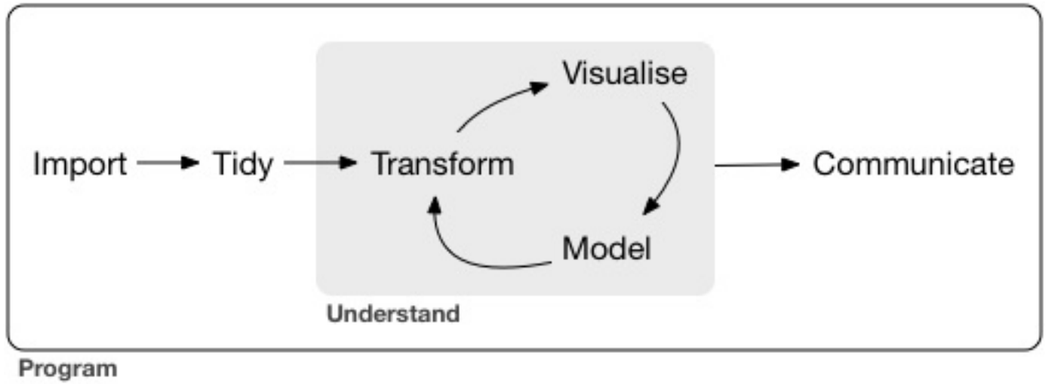
INTRODUCTION TO THE TIDYVERSE

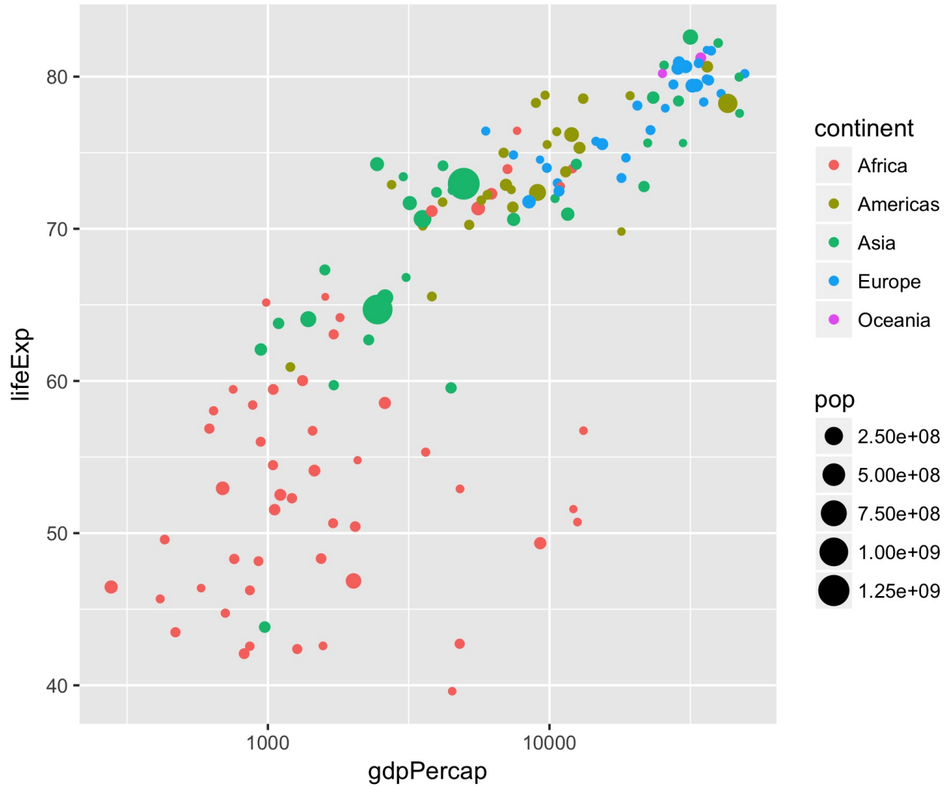
The gapminder dataset

David Robinson
Data Scientist, Stack Overflow

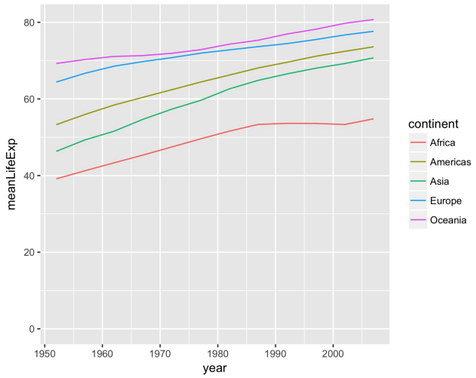
Tidyverse


Visualise
Import
Tidy
Transform
Communicate
Model
Understand
Program

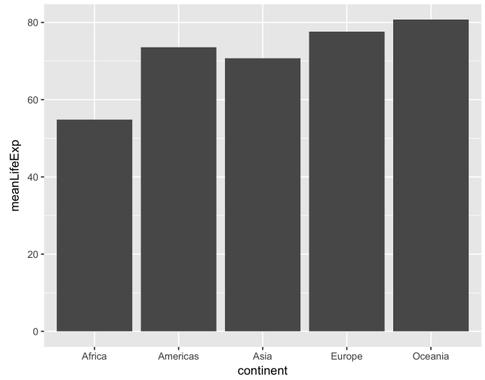
Gapminder


80
70
60
50
40
lifeExp
1000
10000
gdpPercap
continent
Africa
Americas
Asia
Europe
Oceania
pop
2.50e+08
5.00e+08
7.50e+08
1.00e+09
1.25e+09

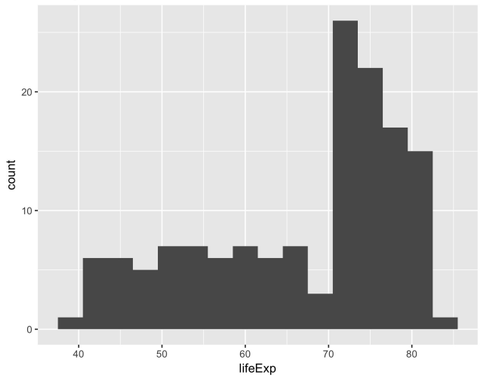
Line plot



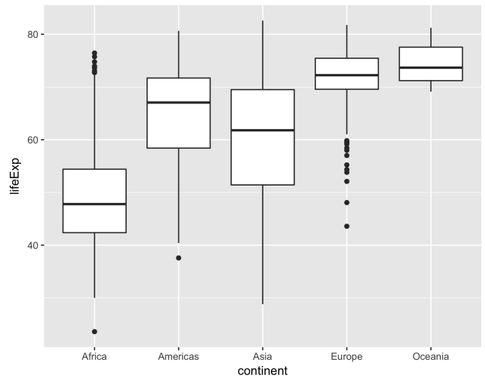
Bar plot



Histogram



Box plot



Loading packages

```
library(gapminder)
```

```
library(dplyr)
```

The gapminder dataset

```
gapminder
```

```
# A tibble: 1,704 x 6
       country continent  year lifeExp      pop gdpPercap
        <fctr>    <fctr> <int>   <dbl>    <dbl>     <dbl>
 1 Afghanistan      Asia  1952  28.801  8425333  779.4453
 2 Afghanistan      Asia  1957  30.332  9240934  820.8530
 3 Afghanistan      Asia  1962  31.997 10267083  853.1007
 4 Afghanistan      Asia  1967  34.020 11537966  836.1971
 5 Afghanistan      Asia  1972  36.088 13079460  739.9811
 6 Afghanistan      Asia  1977  38.438 14880372  786.1134
 7 Afghanistan      Asia  1982  39.854 12881816  978.0114
 8 Afghanistan      Asia  1987  40.822 13867957  852.3959
 9 Afghanistan      Asia  1992  41.674 16317921  649.3414
10 Afghanistan      Asia  1997  41.763 22227415  635.3414
# ... with 1,694 more rows
```

INTRODUCTION TO THE TIDYVERSE

Let's practice!

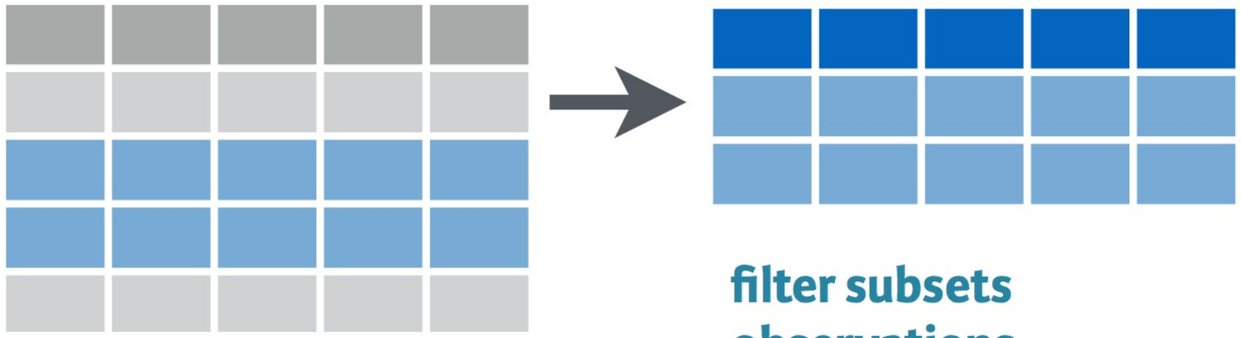
INTRODUCTION TO THE TIDYVERSE

The filter verb

David Robinson
Data Scientist, Stack Overflow

The filter verb

`filter()`



filter subsets observations

Filtering for one year

```
gapminder %>%
  filter(year == 2007)
```

```
# A tibble: 142 x 6
       country continent  year lifeExp       pop  gdpPercap
        <fctr>    <fctr> <int>   <dbl>     <dbl>      <dbl>
 1 Afghanistan      Asia  2007  43.828  31889923   974.5803
 2     Albania    Europe  2007  76.423   3600523  5937.0295
 3     Algeria    Africa  2007  72.301  33333216  6223.3675
 4      Angola    Africa  2007  42.731  12420476  4797.2313
 5   Argentina  Americas  2007  75.320  40301927 12779.3796
 6   Australia   Oceania  2007  81.235  20434176 34435.3674
 7     Austria    Europe  2007  79.829   8199783 36126.4927
 8     Bahrain      Asia  2007  75.635    708573 29796.0483
 9  Bangladesh      Asia  2007  64.062 150448339  1391.2538
10     Belgium    Europe  2007  79.441  10392226 33692.6051
# ... with 132 more rows
```

Filtering for one country

```
gapminder %>%
  filter(country == "United States")
```

```
# A tibble: 12 x 6
         country continent  year lifeExp       pop gdpPercap
          <fctr>    <fctr> <int>   <dbl>     <dbl>     <dbl>
 1 United States  Americas  1952  68.440 157553000  13990.48
 2 United States  Americas  1957  69.490 171984000  14847.13
 3 United States  Americas  1962  70.210 186538000  16173.15
 4 United States  Americas  1967  70.760 198712000  19530.37
 5 United States  Americas  1972  71.340 209896000  21806.04
 6 United States  Americas  1977  73.380 220239000  24072.63
 7 United States  Americas  1982  74.650 232187835  25009.56
 8 United States  Americas  1987  75.020 242803533  29884.35
 9 United States  Americas  1992  76.090 256894189  32003.93
10 United States  Americas  1997  76.810 272911760  35767.43
11 United States  Americas  2002  77.310 287675526  39097.10
12 United States  Americas  2007  78.242 301139947  42951.65
```

Filtering for two variables

```
gapminder %>%
  filter(year == 2007, country == "United States")
```

```
# A tibble: 1 x 6
         country continent  year lifeExp       pop gdpPercap
          <fctr>    <fctr> <int>   <dbl>     <dbl>     <dbl>
1 United States  Americas  2007  78.242 301139947  42951.65
```

INTRODUCTION TO THE TIDYVERSE

Let's practice!

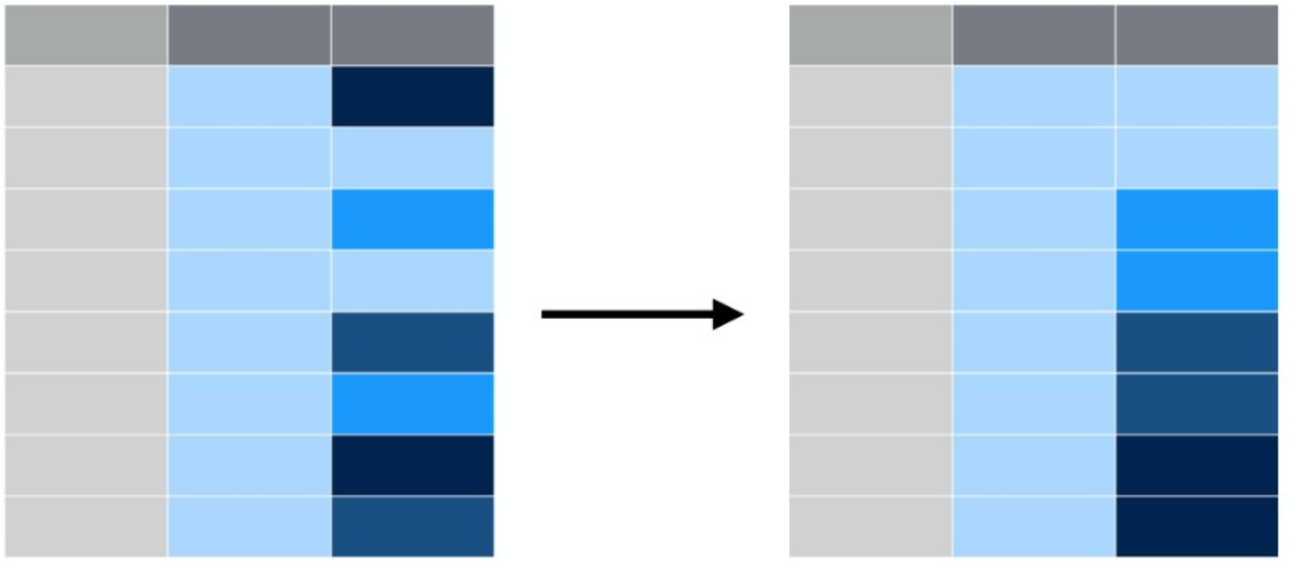
INTRODUCTION TO THE TIDYVERSE

The arrange verb

David Robinson
Data Scientist, Stack Overflow

The arrange verb

`arrange()` sorts a table based on a variable



Sorting with arrange

```
gapminder %>%
  arrange(gdpPercap)
```

```
# A tibble: 1,704 x 6
             country continent  year lifeExp      pop gdpPercap
              <fctr>    <fctr> <int>   <dbl>    <dbl>     <dbl>
 1 Congo, Dem. Rep.    Africa  2002  44.966 55379852  241.1659
 2 Congo, Dem. Rep.    Africa  2007  46.462 64606759  277.5519
 3          Lesotho    Africa  1952  42.138   748747  298.8462
 4    Guinea-Bissau    Africa  1952  32.500   580653  299.8503
 5 Congo, Dem. Rep.    Africa  1997  42.587 47798986  312.1884
 6          Eritrea    Africa  1952  35.928  1438760  328.9406
 7          Myanmar      Asia  1952  36.319 20092996  331.0000
 8          Lesotho    Africa  1957  45.047   813338  335.9971
 9          Burundi    Africa  1952  39.031  2445618  339.2965
10          Eritrea    Africa  1957  38.047  1542611  344.1619
# ... with 1,694 more rows
```

Sorting in descending order

```
gapminder %>%
  arrange(desc(gdpPercap))
```

```
# A tibble: 1,704 x 6
     country continent  year lifeExp     pop gdpPercap
      <fctr>    <fctr> <int>   <dbl>   <dbl>     <dbl>
 1    Kuwait      Asia  1957  58.033  212846 113523.13
 2    Kuwait      Asia  1972  67.712  841934 109347.87
 3    Kuwait      Asia  1952  55.565  160000 108382.35
 4    Kuwait      Asia  1962  60.470  358266  95458.11
 5    Kuwait      Asia  1967  64.624  575003  80894.88
 6    Kuwait      Asia  1977  69.343 1140357  59265.48
 7    Norway    Europe  2007  80.196 4627926  49357.19
 8    Kuwait      Asia  2007  77.588 2505559  47306.99
 9 Singapore      Asia  2007  79.972 4553009  47143.18
10    Norway    Europe  2002  79.050 4535591  44683.98
# ... with 1,694 more rows
```

Filtering then arranging

```
gapminder %>%
  filter(year == 2007) %>%
  arrange(desc(gdpPercap))
```

```
# A tibble: 142 x 6
            country continent  year lifeExp       pop gdpPercap
             <fctr>    <fctr> <int>   <dbl>     <dbl>     <dbl>
 1           Norway    Europe  2007  80.196   4627926  49357.19
 2           Kuwait      Asia  2007  77.588   2505559  47306.99
 3        Singapore      Asia  2007  79.972   4553009  47143.18
 4    United States  Americas  2007  78.242 301139947  42951.65
 5          Ireland    Europe  2007  78.885   4109086  40676.00
 6 Hong Kong, China      Asia  2007  82.208   6980412  39724.98
 7      Switzerland    Europe  2007  81.701   7554661  37506.42
 8      Netherlands    Europe  2007  79.762  16570613  36797.93
 9           Canada  Americas  2007  80.653  33390141  36319.24
10          Iceland    Europe  2007  81.757    301931  36180.79
# ... with 132 more rows
```

INTRODUCTION TO THE TIDYVERSE

Let's practice!

INTRODUCTION TO THE TIDYVERSE

The mutate verb

David Robinson
Data Scientist, Stack Overflow

The mutate verb

`mutate()`

mutate changes or adds variables

Using mutate to change a variable

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

```
# A tibble: 1,704 x 6
       country continent  year lifeExp       pop gdpPercap
        <fctr>    <fctr> <int>   <dbl>     <dbl>     <dbl>
 1 Afghanistan      Asia  1952  28.801  8.425333  779.4453
 2 Afghanistan      Asia  1957  30.332  9.240934  820.8530
 3 Afghanistan      Asia  1962  31.997 10.267083  853.1007
 4 Afghanistan      Asia  1967  34.020 11.537966  836.1971
 5 Afghanistan      Asia  1972  36.088 13.079460  739.9811
 6 Afghanistan      Asia  1977  38.438 14.880372  786.1134
 7 Afghanistan      Asia  1982  39.854 12.881816  978.0114
 8 Afghanistan      Asia  1987  40.822 13.867957  852.3959
 9 Afghanistan      Asia  1992  41.674 16.317921  649.3414
10 Afghanistan      Asia  1997  41.763 22.227415  635.3414
# ... with 1,694 more rows
```

Using mutate to add a new variable

```
gapminder %>%
  mutate(gdp = gdpPercap * pop)
```

```
# A tibble: 1,704 x 7
       country continent  year lifeExp      pop gdpPercap         gdp
        <fctr>    <fctr> <int>   <dbl>    <dbl>     <dbl>       <dbl>
 1 Afghanistan      Asia  1952  28.801  8425333  779.4453  6567086330
 2 Afghanistan      Asia  1957  30.332  9240934  820.8530  7585448670
 3 Afghanistan      Asia  1962  31.997 10267083  853.1007  8758855797
 4 Afghanistan      Asia  1967  34.020 11537966  836.1971  9648014150
 5 Afghanistan      Asia  1972  36.088 13079460  739.9811  9678553274
 6 Afghanistan      Asia  1977  38.438 14880372  786.1134 11697659231
 7 Afghanistan      Asia  1982  39.854 12881816  978.0114 12598563401
 8 Afghanistan      Asia  1987  40.822 13867957  852.3959 11820990309
 9 Afghanistan      Asia  1992  41.674 16317921  649.3414 10595901589
10 Afghanistan      Asia  1997  41.763 22227415  635.3414 14121995875
# ... with 1,694 more rows
```

Combining verbs

```
gapminder %>%
  mutate(gdp = gdpPercap * pop) %>%
  filter(year == 2007) %>%
  arrange(desc(gdp))
```

```
# A tibble: 142 x 7
          country continent  year lifeExp        pop gdpPercap          gdp
           <fctr>    <fctr> <int>   <dbl>      <dbl>     <dbl>        <dbl>
 1  United States  Americas  2007  78.242  301139947 42951.653 1.293446e+13
 2          China      Asia  2007  72.961 1318683096  4959.115 6.539501e+12
 3          Japan      Asia  2007  82.603  127467972 31656.068 4.035135e+12
 4          India      Asia  2007  64.698 1110396331  2452.210 2.722925e+12
 5        Germany    Europe  2007  79.406   82400996 32170.374 2.650871e+12
 6 United Kingdom    Europe  2007  79.425   60776238 33203.261 2.017969e+12
 7         France    Europe  2007  80.657   61083916 30470.017 1.861228e+12
 8         Brazil  Americas  2007  72.390  190010647  9065.801 1.722599e+12
 9          Italy    Europe  2007  80.546   58147733 28569.720 1.661264e+12
10         Mexico  Americas  2007  76.195  108700891 11977.575 1.301973e+12
# ... with 132 more rows
```

INTRODUCTION TO THE TIDYVERSE

Let's practice!