The summarize verb

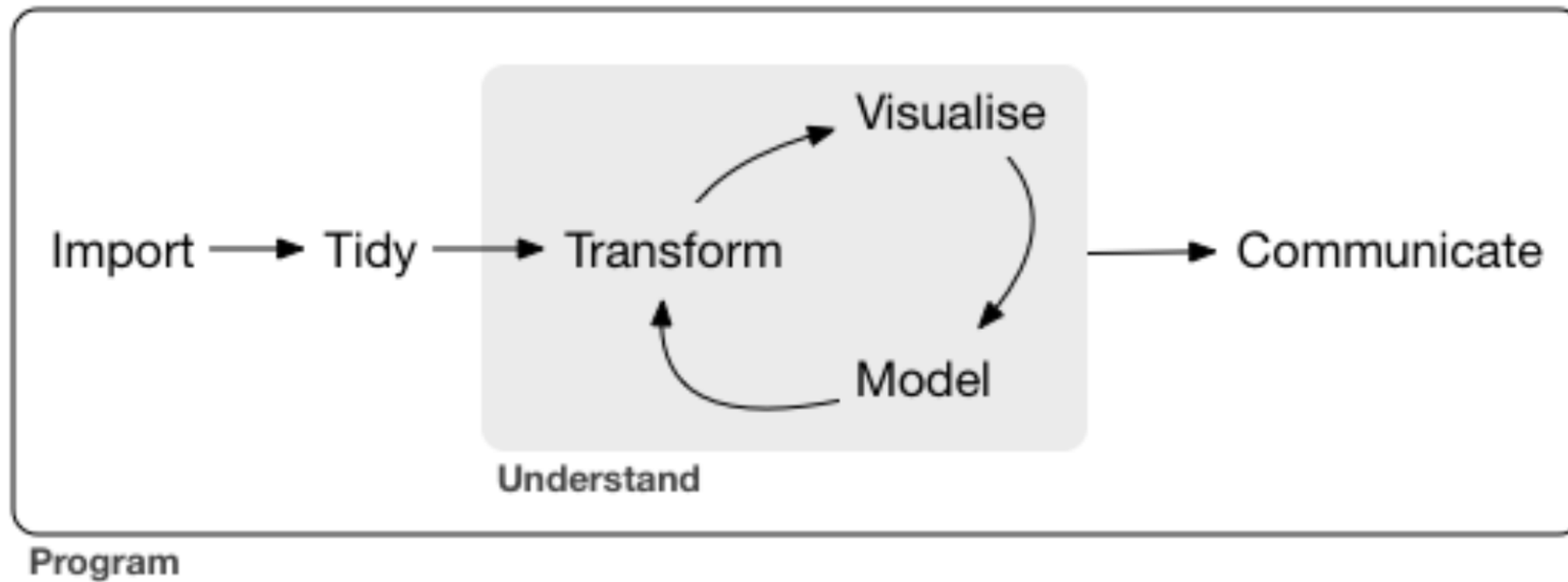## INTRODUCTION TO THE TIDYVERSE

David Robinson
Chief Data Scientist, DataCamp

Data transformation and visualization

Import → Tidy → Transform

Visualise

Model

Understand

Communicate

Program

Extracting data

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

```
# 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
```

The summarize verb

**summarize() turns
many rows into one**


```
gapminder %>%
  summarize(meanLifeExp = mean(lifeExp))
```

```
# A tibble: 1 x 1
  meanLifeExp
        <dbl>
1    59.47444
```

Summarizing one year

```
gapminder %>%
  filter(year == 2007) %>%
  summarize(meanLifeExp = mean(lifeExp))
```

```
# A tibble: 1 x 1
  meanLifeExp
        <dbl>
1    67.00742
```

Summarizing into multiple columns

```r
gapminder %>%
  filter(year == 2007) %>%
  summarize(meanLifeExp = mean(lifeExp),
            totalPop = sum(pop))
```

```
# A tibble: 1 x 2
  meanLifeExp    totalPop
        <dbl>       <dbl>
1    67.00742  6251013179
```

Functions you can use for summarizing

- `mean`

- `sum`

- `median`

- `min`

- `max`

Let's practice!

INTRODUCTION TO THE TIDYVERSE

datacamp

The group_by verb

INTRODUCTION TO THE TIDYVERSE

David Robinson
Chief Data Scientist, DataCamp

The summarize verb

```
gapminder %>%
  filter(year == 2007) %>%
  summarize(meanLifeExp = mean(lifeExp),
            totalPop = sum(pop))
```

```
# A tibble: 1 x 2
  meanLifeExp    totalPop
        <dbl>       <dbl>
1    67.00742 6251013179
```

group_by() before
summarize() turns groups
into one row each

Summarizing by year

```r
gapminder %>%
  group_by(year) %>%
  summarize(meanLifeExp = mean(lifeExp),
            totalPop = sum(pop))
```

```
# A tibble: 12 x 3
    year meanLifeExp    totalPop
   <int>       <dbl>       <dbl>
 1  1952    49.05762  2406957150
 2  1957    51.50740  2664404580
 3  1962    53.60925  2899782974
 4  1967    55.67829  3217478384
 5  1972    57.64739  3576977158
 6  1977    59.57016  3930045807
 7  1982    61.53320  4289436840
 8  1987    63.21261  4691477418
 9  1992    64.16034  5110710260
10  1997    65.01468  5515204472
11  2002    65.69492  5886977579
12  2007    67.00742  6251013179
```

Summarizing by continent

```
gapminder %>%
  filter(year == 2007) %>%
  group_by(continent) %>%
  summarize(meanLifeExp = mean(lifeExp),
            totalPop = sum(pop))
```

```
# A tibble: 5 x 3
  continent meanLifeExp    totalPop
    <fctr>        <dbl>       <dbl>
1    Africa     48.86533  6187585961
2  Americas     64.65874  7351438499
3      Asia     60.06490 30507333901
4    Europe     71.90369  6181115304
5   Oceania     74.32621   212992136
```

Summarizing by continent and year

```
gapminder %>%
  group_by(year, continent) %>%
  summarize(totalPop = sum(pop),
            meanLifeExp = mean(lifeExp))
```

```
# A tibble: 60 x 4
# Groups:   year [?]
    year continent    totalPop meanLifeExp
   <int>    <fctr>       <dbl>       <dbl>
 1  1952    Africa   237640501    39.13550
 2  1952  Americas   345152446    53.27984
 3  1952      Asia  1395357351    46.31439
 4  1952    Europe   418120846    64.40850
 5  1952   Oceania    10686006    69.25500
 6  1957    Africa   264837738    41.26635
 7  1957  Americas   386953916    55.96028
 8  1957      Asia  1562780599    49.31854
 9  1957    Europe   437890351    66.70307
10  1957   Oceania    11941976    70.29500
# ... with 50 more rows
```

Let's practice!

INTRODUCTION TO THE TIDYVERSE

datacamp

Visualizing summarized data

INTRODUCTION TO THE TIDYVERSE

David Robinson
Chief Data Scientist, DataCamp
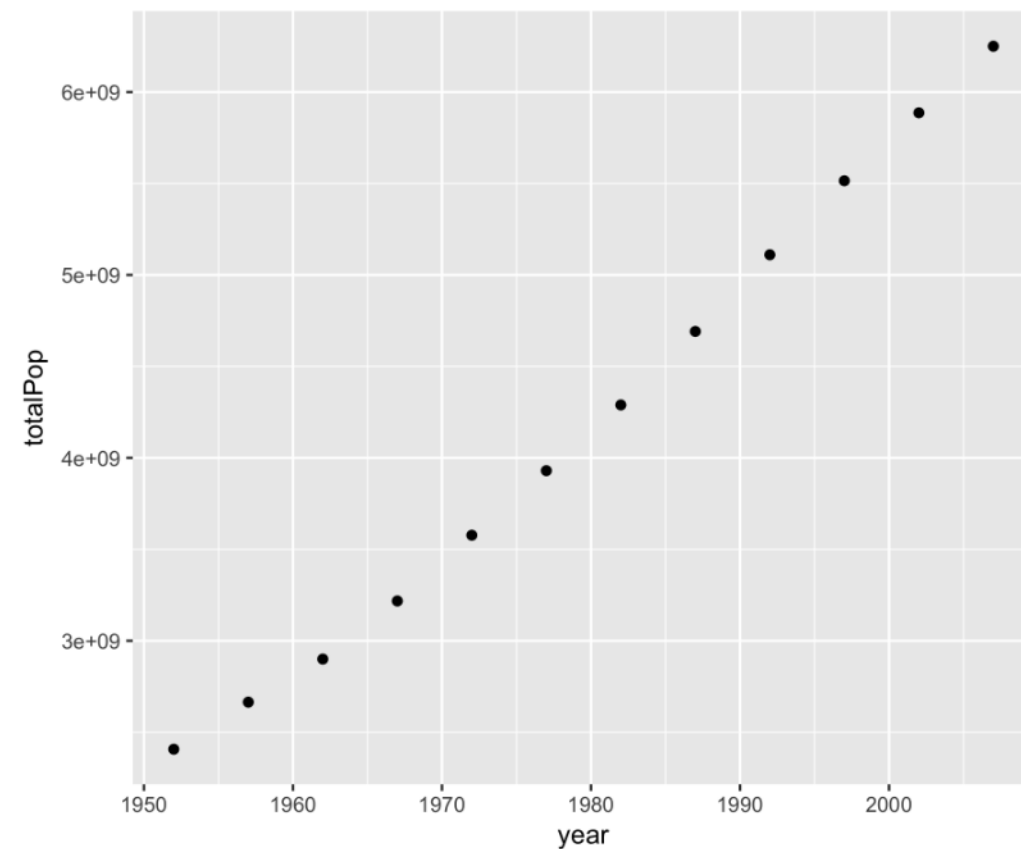
Summarizing by year

```r
by_year <- gapminder %>%
  group_by(year) %>%
  summarize(totalPop = sum(pop),
            meanLifeExp = mean(lifeExp))

by_year
```

```
# A tibble: 12 x 3
    year    totalPop meanLifeExp
   <int>       <dbl>       <dbl>
 1  1952  2406957150    49.05762
 2  1957  2664404580    51.50740
 3  1962  2899782974    53.60925
 4  1967  3217478384    55.67829
 5  1972  3576977158    57.64739
 6  1977  3930045807    59.57016
 7  1982  4289436840    61.53320
 8  1987  4691477418    63.21261
 9  1992  5110710260    64.16034
10  1997  5515204472    65.01468
11  2002  5886977579    65.69492
12  2007  6251013179    67.00742
```
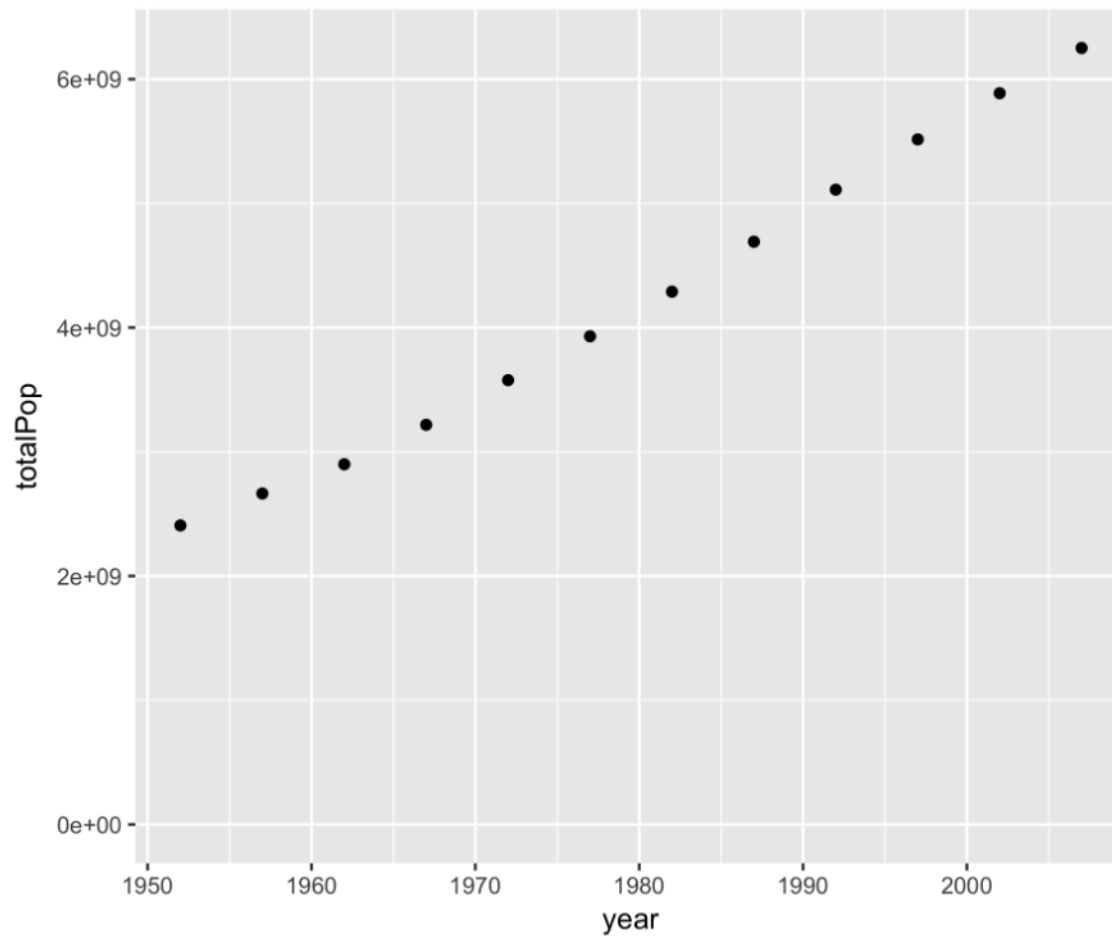
Visualizing population over time

```
ggplot(by_year, aes(x = year, y = totalPop)) +
  geom_point()
```

Starting y-axis at zero

```
ggplot(by_year, aes(x = year, y = totalPop)) +
  geom_point() +
  expand_limits(y = 0)
```
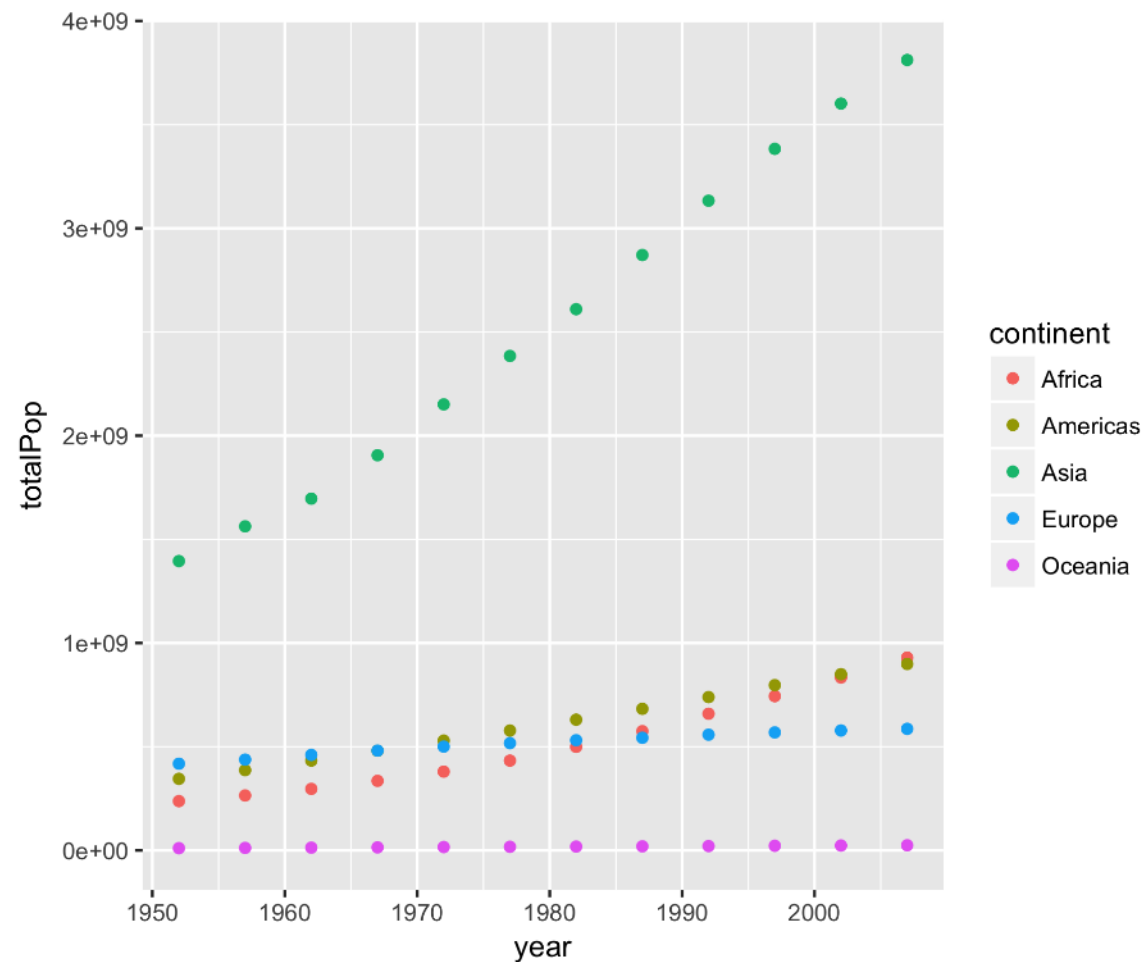
Summarizing by year and continent

```r
by_year_continent <- gapminder %>%
  group_by(year, continent) %>%
  summarize(totalPop = sum(pop),
            meanLifeExp = mean(lifeExp))

by_year_continent
```

```
# A tibble: 60 x 4
# Groups:   year [?]
    year continent    totalPop meanLifeExp
   <int>    <fctr>       <dbl>       <dbl>
 1  1952    Africa   237640501    39.13550
 2  1952  Americas   345152446    53.27984
 3  1952      Asia  1395357351    46.31439
 4  1952    Europe   418120846    64.40850
 5  1952   Oceania    10686006    69.25500
 6  1957    Africa   264837738    41.26635
 7  1957  Americas   386953916    55.96028
 8  1957      Asia  1562780599    49.31854
 9  1957    Europe   437890351    66.70307
10  1957   Oceania    11941976    70.29500
# ... with 50 more rows
```

Visualizing population by year and continent

```
ggplot(by_year_continent, aes(x = year, y = totalPop, color = continent)) +
  geom_point() +
  expand_limits(y = 0)
```

Let's practice!

datacamp