DataCamp: Intro to Tidyverse

N. Kim

2025-06-06 15:35

https://app.datacamp.com/learn/courses/introduction-to-the-tidyverse

Notes started on 2025-06-05.

filter(), arrange(), mutate()

Comparison chart

Task	SAS	R Tidyverse %>% clause
Filter rows	`where` crit1 `and` crit2;	`filter`(crit1, crit2)
	`where` crit1 `or` crit2;	`filter`(crit1 | crit2)
Sort rows	`proc sort`; `by` var1 `descending` var2;	`arrange(var1, desc`(var2))
Transform an existing variable	var = var / 100;	`mutate`(var = var / 100)
Add a new variable	newvar = var1 * 100;	`mutate`(newvar = var1 * 100)

ggplot2

`ggplot`(dataset, `aes(x=`ivar, `y=`dvar, ...)) + [...]

`aes` clause

Type	ggplot2
`x=`	variable for x-axis
`y=`	variable for y-axis
`color=`	group variable for color coding
`size=`	group variable for size aesthestic

Additional clauses

+ clause	result
`geom_point()`	produce a scatterplot
`geom_boxplot()`	produce a boxplot
`geom_line()`	produce a line plot (not necessarily a straight line)
`geom_col()`	produce a bar plot (note: all bar plots start at y=0)
`geom_histogram()`	produce a histogram* (bins vs counts)
`geom_histogram(binwidth=5)`	`binwidth=` # units bundled together for each bar
`geom_histogram(bins=20)`	`bins=` # bars
`geom_boxplot()`	produce a boxplot
`scale_x_log10()`	have x-axis on a log scale
`scale_y_log10()`	have y-axis on a log scale
`facet_wrap(~ groupVar)`	produce subplots, one for each value of the grouping variable
`expand_limits(y = 0)`	start the y-axis at zero
`labs(title=, x=, y=)`	Add plot title and/or labels for x- and y- axes.

*A histogram can have an x= or a y= aesthestic, but not both.

summarize()

`+ summarize`(newVar = function(existingVar))

Functions you can use for summarizing:

- `sum()`
- `mean()`, `median()`
- `min()`, `max()`

group_by()

Combine group_by() with summarize() to get stats for each group (or combination of groups).

```
library(dplyr)
library(ggplot2)
library(gapminder)
```

```r
head(gapminder)
```

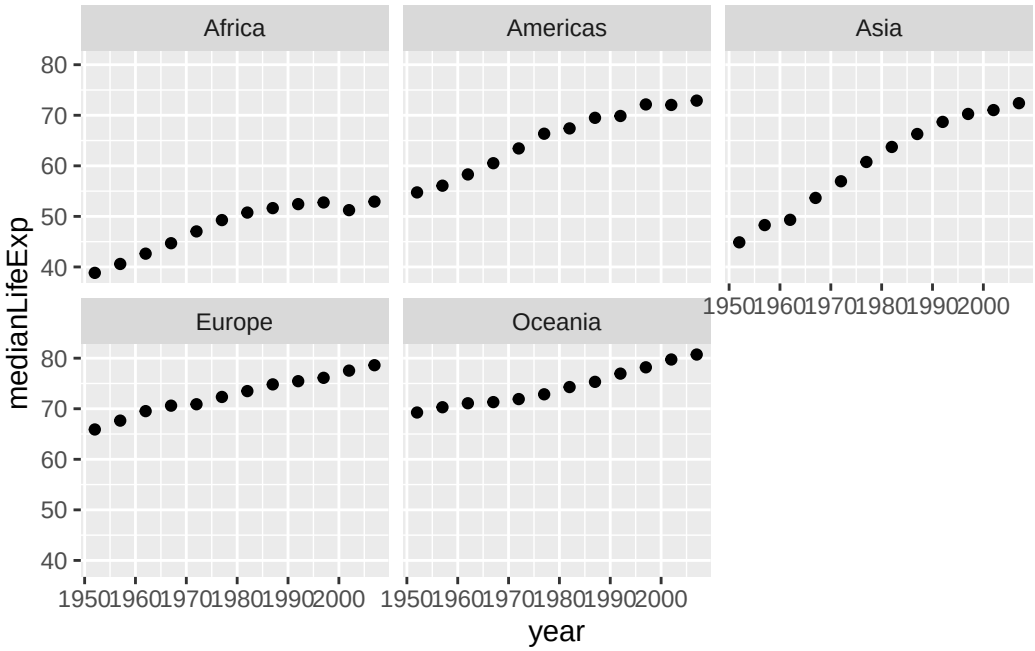
```
# A tibble: 6 x 6
  country     continent  year lifeExp      pop gdpPercap
  <fct>       <fct>     <int>   <dbl>    <int>     <dbl>
1 Afghanistan Asia       1952    28.8  8425333      779.
2 Afghanistan Asia       1957    30.3  9240934      821.
3 Afghanistan Asia       1962    32.0 10267083      853.
4 Afghanistan Asia       1967    34.0 11537966      836.
5 Afghanistan Asia       1972    36.1 13079460      740.
6 Afghanistan Asia       1977    38.4 14880372      786.
```

```r
# Find median life expectancy, by continent and year
median_by_continent_year <- gapminder %>%
    group_by(continent, year) %>%
    summarize(medianLifeExp = median(lifeExp))
```

```r
median_by_continent_year
```

```
# A tibble: 60 x 3
# Groups:   continent [5]
   continent  year medianLifeExp
   <fct>     <int>         <dbl>
 1 Africa     1952          38.8
 2 Africa     1957          40.6
 3 Africa     1962          42.6
 4 Africa     1967          44.7
 5 Africa     1972          47.0
 6 Africa     1977          49.3
 7 Africa     1982          50.8
 8 Africa     1987          51.6
 9 Africa     1992          52.4
10 Africa     1997          52.8
# i 50 more rows
```

```r
ggplot(median_by_continent_year, aes(x=year, y=medianLifeExp)) +
  geom_point() +
  facet_wrap(~ continent)
```


Africa
Americas
Asia
Europe
Oceania
medianLifeExp
80
70
60
50
40
1950
1960
1970
1980
1990
2000
year