

DataCamp: Intro to Tidyverse

N. Kim

2025-06-05 16:47

<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	<code>where crit1 and crit2;</code> <code>where crit1 or crit2;</code>	<code>filter(crit1, crit2)</code> <code>filter(crit1 crit2)</code>
Sort rows	<code>proc sort;</code> <code>by var1 descending var2;</code>	<code>arrange(var1, desc(var2))</code>
Transform an existing variable	<code>var = var / 100;</code>	<code>mutate(var = var / 100)</code>
Add a new variable	<code>newvar = var1 * 100;</code>	<code>mutate(newvar = var1 * 100)</code>

ggplot2

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

aes clause

Type	ggplot2
<code>x=</code>	variable for x-axis
<code>y=</code>	variable for y-axis
<code>color=</code>	group variable for color coding
<code>size=</code>	group variable for size aesthetic

Additional clauses

+ clause	result
<code>geom_point()</code>	produce a scatterplot
<code>geom_boxplot()</code>	produce a boxplot
<code>scale_x_log10()</code>	have x-axis on a log scale
<code>scale_y_log10()</code>	have y-axis on a log scale
<code>facet_wrap(~ groupVar)</code>	produce subplots, one for each value of the grouping variable
<code>expand_limits(y = 0)</code>	start the y-axis at zero

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)
```

```
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.
```

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

```
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
```

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

