summarize

made with flipbookr and xaringan

Gina Reynolds, January 2020

```
library(gapminder)
library(tidyverse)
```

```
── Attaching packages ─────────────────────────────────────── tidyverse 1.3.0 ──

✓ ggplot2 3.3.2     ✓ purrr   0.3.3
✓ tibble  3.0.0     ✓ dplyr   0.8.5
✓ tidyr   1.0.2     ✓ stringr 1.4.0
✓ readr   1.3.1     ✓ forcats 0.5.0

Warning: package 'ggplot2' was built under R version 3.6.2

Warning: package 'tibble' was built under R version 3.6.2

── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
```

```
gapminder
```

```
# A tibble: 1,704 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.
 7 Afghanistan Asia       1982    39.9 12881816      978.
 8 Afghanistan Asia       1987    40.8 13867957      852.
 9 Afghanistan Asia       1992    41.7 16317921      649.
10 Afghanistan Asia       1997    41.8 22227415      635.
# … with 1,694 more rows
```

```
gapminder %>%
  mutate(num_rows = n())
```

```
# A tibble: 1,704 x 7
   country     continent  year lifeExp      pop gdpPercap num_rows
   <fct>       <fct>     <int>   <dbl>    <int>     <dbl>    <int>
 1 Afghanistan Asia       1952    28.8  8425333      779.     1704
 2 Afghanistan Asia       1957    30.3  9240934      821.     1704
 3 Afghanistan Asia       1962    32.0 10267083      853.     1704
 4 Afghanistan Asia       1967    34.0 11537966      836.     1704
 5 Afghanistan Asia       1972    36.1 13079460      740.     1704
 6 Afghanistan Asia       1977    38.4 14880372      786.     1704
 7 Afghanistan Asia       1982    39.9 12881816      978.     1704
 8 Afghanistan Asia       1987    40.8 13867957      852.     1704
 9 Afghanistan Asia       1992    41.7 16317921      649.     1704
10 Afghanistan Asia       1997    41.8 22227415      635.     1704
# … with 1,694 more rows
```

```
gapminder %>%
  mutate(num_rows = n()) ->
gap_n
```

```
gapminder %>%
  mutate(num_rows = n()) ->
gap_n

# a pipeline to count the number
# of observations by group
gapminder
```

```
# A tibble: 1,704 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.
 7 Afghanistan Asia       1982    39.9 12881816       978.
 8 Afghanistan Asia       1987    40.8 13867957       852.
 9 Afghanistan Asia       1992    41.7 16317921       649.
10 Afghanistan Asia       1997    41.8 22227415       635.
# … with 1,694 more rows
```

```r
gapminder %>%
  mutate(num_rows = n()) ->
gap_n

# a pipeline to count the number
# of observations by group
gapminder %>%
  distinct(country, continent)
```

```
# A tibble: 142 x 2
   country     continent
   <fct>       <fct>
 1 Afghanistan Asia
 2 Albania     Europe
 3 Algeria     Africa
 4 Angola      Africa
 5 Argentina   Americas
 6 Australia   Oceania
 7 Austria     Europe
 8 Bahrain     Asia
 9 Bangladesh  Asia
10 Belgium     Europe
# … with 132 more rows
```

```r
gapminder %>%
  mutate(num_rows = n()) ->
gap_n

# a pipeline to count the number
# of observations by group
gapminder %>%
  distinct(country, continent) %>%
# tally and count are the same
  count(continent)
```

```
# A tibble: 5 x 2
  continent     n
  <fct>     <int>
1 Africa       52
2 Americas     25
3 Asia         33
4 Europe       30
5 Oceania       2
```

```r
gapminder %>%
  mutate(num_rows = n()) ->
gap_n

# a pipeline to count the number
# of observations by group
gapminder %>%
  distinct(country, continent) %>%
# tally and count are the same
  count(continent) %>%
  rename(num_countries = n)
```

```
# A tibble: 5 x 2
  continent num_countries
  <fct>             <int>
1 Africa               52
2 Americas             25
3 Asia                 33
4 Europe               30
5 Oceania               2
```

```
gapminder %>%
  mutate(num_rows = n()) ->
gap_n

# a pipeline to count the number
# of observations by group
gapminder %>%
  distinct(country, continent) %>%
# tally and count are the same
  count(continent) %>%
  rename(num_countries = n) ->
count_in_continents
```

```
gapminder %>%
  mutate(num_rows = n()) ->
gap_n

# a pipeline to count the number
# of observations by group
gapminder %>%
  distinct(country, continent) %>%
# tally and count are the same
  count(continent) %>%
  rename(num_countries = n) ->
count_in_continents

# alternatively use n function
gapminder
```

```
# A tibble: 1,704 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.
 7 Afghanistan Asia       1982    39.9 12881816       978.
 8 Afghanistan Asia       1987    40.8 13867957       852.
 9 Afghanistan Asia       1992    41.7 16317921       649.
10 Afghanistan Asia       1997    41.8 22227415       635.
# … with 1,694 more rows
```

```
gapminder %>%
  mutate(num_rows = n()) ->
gap_n

# a pipeline to count the number
# of observations by group
gapminder %>%
  distinct(country, continent) %>%
# tally and count are the same
  count(continent) %>%
  rename(num_countries = n) ->
count_in_continents

# alternatively use n function
gapminder %>%
  distinct(country, continent)
```

```
# A tibble: 142 x 2
   country     continent
   <fct>       <fct>
 1 Afghanistan Asia
 2 Albania     Europe
 3 Algeria     Africa
 4 Angola      Africa
 5 Argentina   Americas
 6 Australia   Oceania
 7 Austria     Europe
 8 Bahrain     Asia
 9 Bangladesh  Asia
10 Belgium     Europe
# … with 132 more rows
```

```
gapminder %>%
  mutate(num_rows = n()) ->
gap_n

# a pipeline to count the number
# of observations by group
gapminder %>%
  distinct(country, continent) %>%
# tally and count are the same
  count(continent) %>%
  rename(num_countries = n) ->
count_in_continents

# alternatively use n function
gapminder %>%
  distinct(country, continent) %>%
# tally and count are the same
  group_by(continent)
```

```
# A tibble: 142 x 2
# Groups:   continent [5]
   country     continent
   <fct>       <fct>
 1 Afghanistan Asia
 2 Albania     Europe
 3 Algeria     Africa
 4 Angola      Africa
 5 Argentina   Americas
 6 Australia   Oceania
 7 Austria     Europe
 8 Bahrain     Asia
 9 Bangladesh  Asia
10 Belgium     Europe
# … with 132 more rows
```

```
gapminder %>%
  mutate(num_rows = n()) ->
gap_n

# a pipeline to count the number
# of observations by group
gapminder %>%
  distinct(country, continent) %>%
# tally and count are the same
  count(continent) %>%
  rename(num_countries = n) ->
count_in_continents

# alternatively use n function
gapminder %>%
  distinct(country, continent) %>%
# tally and count are the same
  group_by(continent) %>%
  summarise(num_countries = n())
```

```
# A tibble: 5 x 2
  continent num_countries
  <fct>             <int>
1 Africa               52
2 Americas             25
3 Asia                 33
4 Europe               30
5 Oceania               2
```

```
gapminder %>%
  mutate(num_rows = n()) ->
gap_n

# a pipeline to count the number
# of observations by group
gapminder %>%
  distinct(country, continent) %>%
# tally and count are the same
  count(continent) %>%
  rename(num_countries = n) ->
count_in_continents

# alternatively use n function
gapminder %>%
  distinct(country, continent) %>%
# tally and count are the same
  group_by(continent) %>%
  summarise(num_countries = n()) ->
count_in_continents
```

```
# a pipeline to create overall
# variable summaries
gapminder
```

```
# A tibble: 1,704 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.
 7 Afghanistan Asia       1982    39.9 12881816      978.
 8 Afghanistan Asia       1987    40.8 13867957      852.
 9 Afghanistan Asia       1992    41.7 16317921      649.
10 Afghanistan Asia       1997    41.8 22227415      635.
# … with 1,694 more rows
```

```r
# a pipeline to create overall
# variable summaries
gapminder %>%
  filter(year == 2002)
```

```
# A tibble: 142 x 6
   country     continent  year lifeExp       pop gdpPercap
   <fct>       <fct>     <int>   <dbl>     <int>     <dbl>
 1 Afghanistan Asia       2002    42.1  25268405      727.
 2 Albania     Europe     2002    75.7   3508512     4604.
 3 Algeria     Africa     2002    71.0  31287142     5288.
 4 Angola      Africa     2002    41.0  10866106     2773.
 5 Argentina   Americas   2002    74.3  38331121     8798.
 6 Australia   Oceania    2002    80.4  19546792    30688.
 7 Austria     Europe     2002    79.0   8148312    32418.
 8 Bahrain     Asia       2002    74.8    656397    23404.
 9 Bangladesh  Asia       2002    62.0 135656790     1136.
10 Belgium     Europe     2002    78.3  10311970    30486.
# … with 132 more rows
```

```r
# a pipeline to create overall
# variable summaries
gapminder %>%
  filter(year == 2002) %>%
  summarize(mean_life_exp = mean(lifeExp),
            median_life_exp = median(lifeExp))
```

```
# A tibble: 1 x 2
  mean_life_exp median_life_exp
          <dbl>           <dbl>
1          65.7            70.8
```

```r
# a pipeline to create overall
# variable summaries
gapminder %>%
  filter(year == 2002) %>%
  summarize(mean_life_exp = mean(lifeExp),
            median_life_exp = median(lifeExp)) ->
overall_summaries_2002
```

```r
# a pipeline to create overall
# variable summaries
gapminder %>%
  filter(year == 2002) %>%
  summarize(mean_life_exp = mean(lifeExp),
            median_life_exp = median(lifeExp)) ->
overall_summaries_2002

# a pipeline to create groupwise
# variable summaries
gapminder
```

```
# A tibble: 1,704 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.
 7 Afghanistan Asia       1982    39.9 12881816       978.
 8 Afghanistan Asia       1987    40.8 13867957       852.
 9 Afghanistan Asia       1992    41.7 16317921       649.
10 Afghanistan Asia       1997    41.8 22227415       635.
# … with 1,694 more rows
```

```
# a pipeline to create overall
# variable summaries
gapminder %>%
  filter(year == 2002) %>%
  summarize(mean_life_exp = mean(lifeExp),
            median_life_exp = median(lifeExp)) ->
overall_summaries_2002

# a pipeline to create groupwise
# variable summaries
gapminder %>%
  filter(year == 2002)
```

```
# A tibble: 142 x 6
   country     continent  year lifeExp       pop gdpPercap
   <fct>       <fct>     <int>   <dbl>     <int>     <dbl>
 1 Afghanistan Asia       2002    42.1  25268405      727.
 2 Albania     Europe     2002    75.7   3508512     4604.
 3 Algeria     Africa     2002    71.0  31287142     5288.
 4 Angola      Africa     2002    41.0  10866106     2773.
 5 Argentina   Americas   2002    74.3  38331121     8798.
 6 Australia   Oceania    2002    80.4  19546792    30688.
 7 Austria     Europe     2002    79.0   8148312    32418.
 8 Bahrain     Asia       2002    74.8    656397    23404.
 9 Bangladesh  Asia       2002    62.0 135656790     1136.
10 Belgium     Europe     2002    78.3  10311970    30486.
# … with 132 more rows
```

```r
# a pipeline to create overall
# variable summaries
gapminder %>%
  filter(year == 2002) %>%
  summarize(mean_life_exp = mean(lifeExp),
            median_life_exp = median(lifeExp)) ->
overall_summaries_2002

# a pipeline to create groupwise
# variable summaries
gapminder %>%
  filter(year == 2002) %>%
  group_by(continent)
```

```
# A tibble: 142 x 6
# Groups:   continent [5]
   country     continent  year lifeExp       pop gdpPercap
   <fct>       <fct>     <int>   <dbl>     <int>     <dbl>
 1 Afghanistan Asia       2002    42.1  25268405      727.
 2 Albania     Europe     2002    75.7   3508512     4604.
 3 Algeria     Africa     2002    71.0  31287142     5288.
 4 Angola      Africa     2002    41.0  10866106     2773.
 5 Argentina   Americas   2002    74.3  38331121     8798.
 6 Australia   Oceania    2002    80.4  19546792    30688.
 7 Austria     Europe     2002    79.0   8148312    32418.
 8 Bahrain     Asia       2002    74.8    656397    23404.
 9 Bangladesh  Asia       2002    62.0 135656790     1136.
10 Belgium     Europe     2002    78.3  10311970    30486.
# … with 132 more rows
```

```r
# a pipeline to create overall
# variable summaries
gapminder %>%
  filter(year == 2002) %>%
  summarize(mean_life_exp = mean(lifeExp),
            median_life_exp = median(lifeExp)) ->
overall_summaries_2002

# a pipeline to create groupwise
# variable summaries
gapminder %>%
  filter(year == 2002) %>%
  group_by(continent) %>%
  summarize(mean_life_exp = mean(lifeExp),
            median_life_exp = median(lifeExp))
```

```
# A tibble: 5 x 3
  continent mean_life_exp median_life_exp
  <fct>             <dbl>           <dbl>
1 Africa             53.3            51.2
2 Americas           72.4            72.0
3 Asia               69.2            71.0
4 Europe             76.7            77.5
5 Oceania            79.7            79.7
```

```
# a pipeline to create overall
# variable summaries
gapminder %>%
  filter(year == 2002) %>%
  summarize(mean_life_exp = mean(lifeExp),
            median_life_exp = median(lifeExp)) ->
overall_summaries_2002

# a pipeline to create groupwise
# variable summaries
gapminder %>%
  filter(year == 2002) %>%
  group_by(continent) %>%
  summarize(mean_life_exp = mean(lifeExp),
            median_life_exp = median(lifeExp)) ->
summaries_by_continent_2002
```

```r
# a pipeline to create overall
# variable summaries
gapminder %>%
  filter(year == 2002) %>%
  summarize(mean_life_exp = mean(lifeExp),
            median_life_exp = median(lifeExp)) ->
overall_summaries_2002

# a pipeline to create groupwise
# variable summaries
gapminder %>%
  filter(year == 2002) %>%
  group_by(continent) %>%
  summarize(mean_life_exp = mean(lifeExp),
            median_life_exp = median(lifeExp)) ->
summaries_by_continent_2002

# coming soon in a new version of dplyr
# gapminder %>%
#   group_by(continent)   #   group_by(continent) %>%
```

```
# a pipeline to create overall
# variable summaries
gapminder %>%
  filter(year == 2002) %>%
  summarize(mean_life_exp = mean(lifeExp),
            median_life_exp = median(lifeExp)) ->
overall_summaries_2002

# a pipeline to create groupwise
# variable summaries
gapminder %>%
  filter(year == 2002) %>%
  group_by(continent) %>%
  summarize(mean_life_exp = mean(lifeExp),
            median_life_exp = median(lifeExp)) ->
summaries_by_continent_2002

# coming soon in a new version of dplyr
# gapminder %>%
#   group_by(continent) %>%  #   group_by(continent)
#   summarize(across(lifeExp:pop, mean))  #   summari
```