Datacamp_Grouping_and_Summarizing

dizhen

2019/4/4

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

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyverse)
```

```
## -- Attaching packages -------------------------- tidyverse 1.2.1 --

## v tibble  2.1.3     v purrr   0.3.2
## v tidyr   0.8.3     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0

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

```
gapminder <- read.table(file = 'data/gapminder.tsv', sep = '\t', header = TRUE)
```

```
# Extracting data
gapminder %>% filter(country == "United States", year == 2007) %>% head()
```

```
##         country continent year lifeExp       pop gdpPercap
## 1 United States  Americas 2007  78.242 301139947  42951.65
```

The summarize verb

summarize() turns many rows into one

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

```
##   meanLifeExp
## 1    59.51495
```

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

```
##   meanLifeExp
## 1    67.00742
```

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

```
##   meanLifeExp   totalPop
## 1    67.00742 6251013179
```

Functions you can use for sumarizing:

- mean, sum, median, min, max

```r
# Summarize to find the median life expectancy
gapminder %>%
  summarize(medianLifeExp = median(lifeExp))
```

```
##   medianLifeExp
## 1        60.808
```

```r
# Filter for 1957 then summarize the median life expectancy
gapminder %>%
  filter(year == 1957) %>%
  summarize(medianLifeExp = median(lifeExp))
```

```
##   medianLifeExp
## 1       48.3605
```

```r
# Filter for 1957 then summarize the median life expectancy and the maximum GDP per capita
gapminder %>% filter(year == 1957) %>% summarize(medianLifeExp = median(lifeExp),maxGdpPercap = max(gdpP
```

```
##   medianLifeExp maxGdpPercap
## 1       48.3605     113523.1
```

The group_by verb

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

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

```
##   meanLifeExp   totalPop
## 1    67.00742 6251013179
```

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

```
## Warning in summarise_impl(.data, dots, environment(), caller_env()):
## integer overflow - use sum(as.numeric(.))

## Warning in summarise_impl(.data, dots, environment(), caller_env()):
## integer overflow - use sum(as.numeric(.))

## Warning in summarise_impl(.data, dots, environment(), caller_env()):
## integer overflow - use sum(as.numeric(.))

## Warning in summarise_impl(.data, dots, environment(), caller_env()):
## integer overflow - use sum(as.numeric(.))

## Warning in summarise_impl(.data, dots, environment(), caller_env()):
## integer overflow - use sum(as.numeric(.))

## Warning in summarise_impl(.data, dots, environment(), caller_env()):
## integer overflow - use sum(as.numeric(.))

## Warning in summarise_impl(.data, dots, environment(), caller_env()):
## integer overflow - use sum(as.numeric(.))

## Warning in summarise_impl(.data, dots, environment(), caller_env()):
## integer overflow - use sum(as.numeric(.))

## Warning in summarise_impl(.data, dots, environment(), caller_env()):
## integer overflow - use sum(as.numeric(.))

## Warning in summarise_impl(.data, dots, environment(), caller_env()):
## integer overflow - use sum(as.numeric(.))

## Warning in summarise_impl(.data, dots, environment(), caller_env()):
## integer overflow - use sum(as.numeric(.))

## Warning in summarise_impl(.data, dots, environment(), caller_env()):
## integer overflow - use sum(as.numeric(.))
```

```
## # A tibble: 6 x 3
##    year meanLifeExp totalPop
##   <int>       <dbl>    <int>
## 1  1952        49.1       NA
## 2  1957        51.5       NA
## 3  1962        53.7       NA
## 4  1967        55.7       NA
## 5  1972        57.7       NA
## 6  1977        59.6       NA
```

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

```
## Warning in summarise_impl(.data, dots, environment(), caller_env()):
## integer overflow - use sum(as.numeric(.))
```

```
## # A tibble: 5 x 3
##   continent meanLifeExp  totalPop
##   <fct>           <dbl>     <int>
## 1 Africa           54.8 929539692
## 2 Americas         73.6 898871184
## 3 Asia             70.7        NA
## 4 Europe           77.6 586098529
## 5 Oceania          80.7  24549947
```

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

```
## Warning in summarise_impl(.data, dots, environment(), caller_env()):
## integer overflow - use sum(as.numeric(.))

## Warning in summarise_impl(.data, dots, environment(), caller_env()):
## integer overflow - use sum(as.numeric(.))

## Warning in summarise_impl(.data, dots, environment(), caller_env()):
## integer overflow - use sum(as.numeric(.))

## Warning in summarise_impl(.data, dots, environment(), caller_env()):
## integer overflow - use sum(as.numeric(.))

## Warning in summarise_impl(.data, dots, environment(), caller_env()):
## integer overflow - use sum(as.numeric(.))

## Warning in summarise_impl(.data, dots, environment(), caller_env()):
## integer overflow - use sum(as.numeric(.))

## Warning in summarise_impl(.data, dots, environment(), caller_env()):
## integer overflow - use sum(as.numeric(.))

## Warning in summarise_impl(.data, dots, environment(), caller_env()):
## integer overflow - use sum(as.numeric(.))
```

```
## # A tibble: 6 x 4
## # Groups:   year [2]
##    year continent  totalPop meanLifeExp
##   <int> <fct>         <int>       <dbl>
## 1  1952 Africa    234663482        39.1
## 2  1952 Americas  345152446        53.3
## 3  1952 Asia      1395357351        46.3
## 4  1952 Europe     418120846        64.4
## 5  1952 Oceania     10686006        69.3
## 6  1957 Africa     264837738        41.3
```

Practice

```
# Find median life expectancy and maximum GDP per capita in each year
gapminder %>%
  group_by(year) %>%
  summarize(medianLifeExp = median(lifeExp),
            maxGdpPercap = max(gdpPercap)) %>%head()
```

```
## # A tibble: 6 x 3
##    year medianLifeExp maxGdpPercap
##   <int>         <dbl>        <dbl>
## 1  1952          45.3      108382.
## 2  1957          48.4      113523.
## 3  1962          51.5       95458.
## 4  1967          53.8       80895.
## 5  1972          56.5      109348.
## 6  1977          59.7       59265.
```

Summarizing by continent

```
# Find median life expectancy and maximum GDP per capita in each continent in 1957
gapminder %>%
  filter(year == 1957) %>%
  group_by(continent) %>%
  summarize(medianLifeExp = median(lifeExp),
            maxGdpPercap = max(gdpPercap))
```

```
## # A tibble: 5 x 3
##   continent medianLifeExp maxGdpPercap
##   <fct>             <dbl>        <dbl>
## 1 Africa             40.6        5487.
## 2 Americas           56.1       14847.
## 3 Asia               48.3      113523.
## 4 Europe             67.6       17909.
## 5 Oceania            70.3       12247.
```

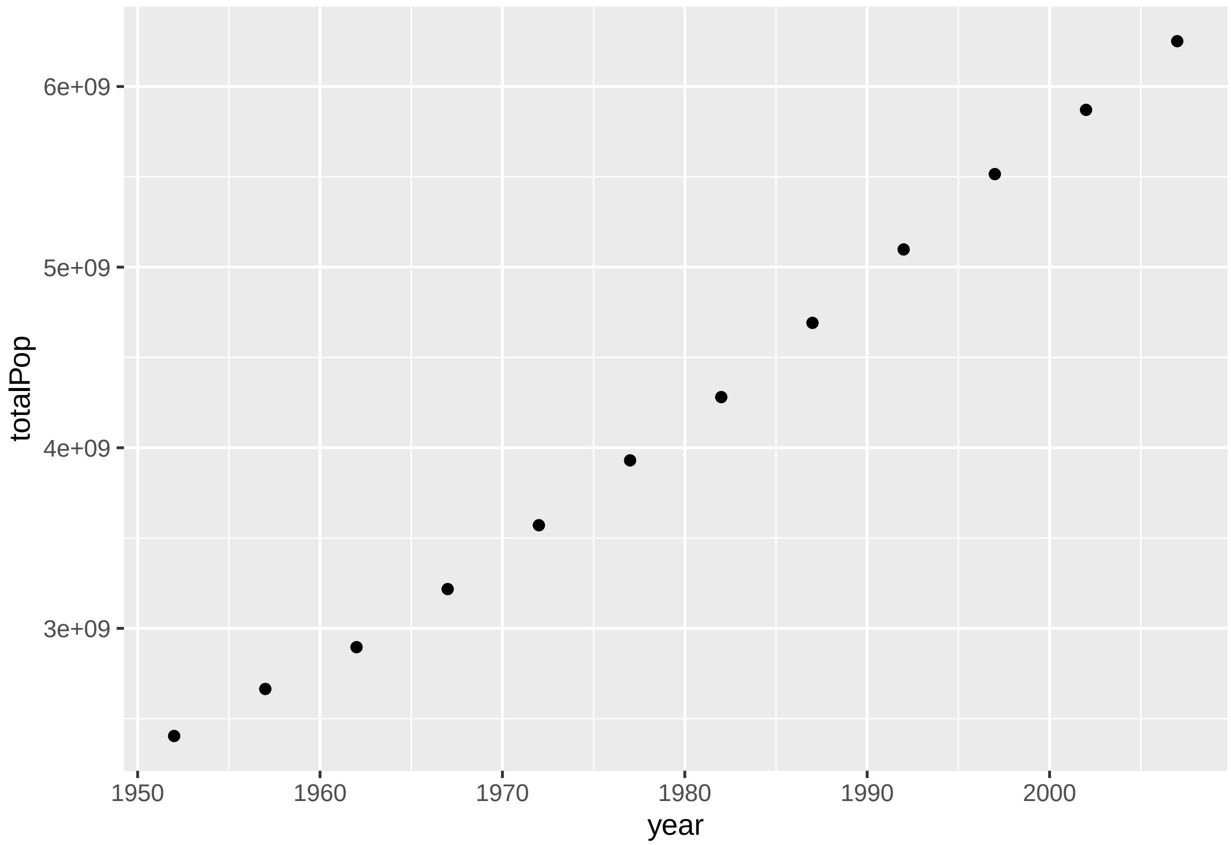
Summarizing by continent and year

```
# Find median life expectancy and maximum GDP per capita in each continent/year combination
gapminder %>%
  group_by(continent, year) %>%
  summarize(medianLifeExp = median(lifeExp),
            maxGdpPercap = max(gdpPercap)) %>%head()
```

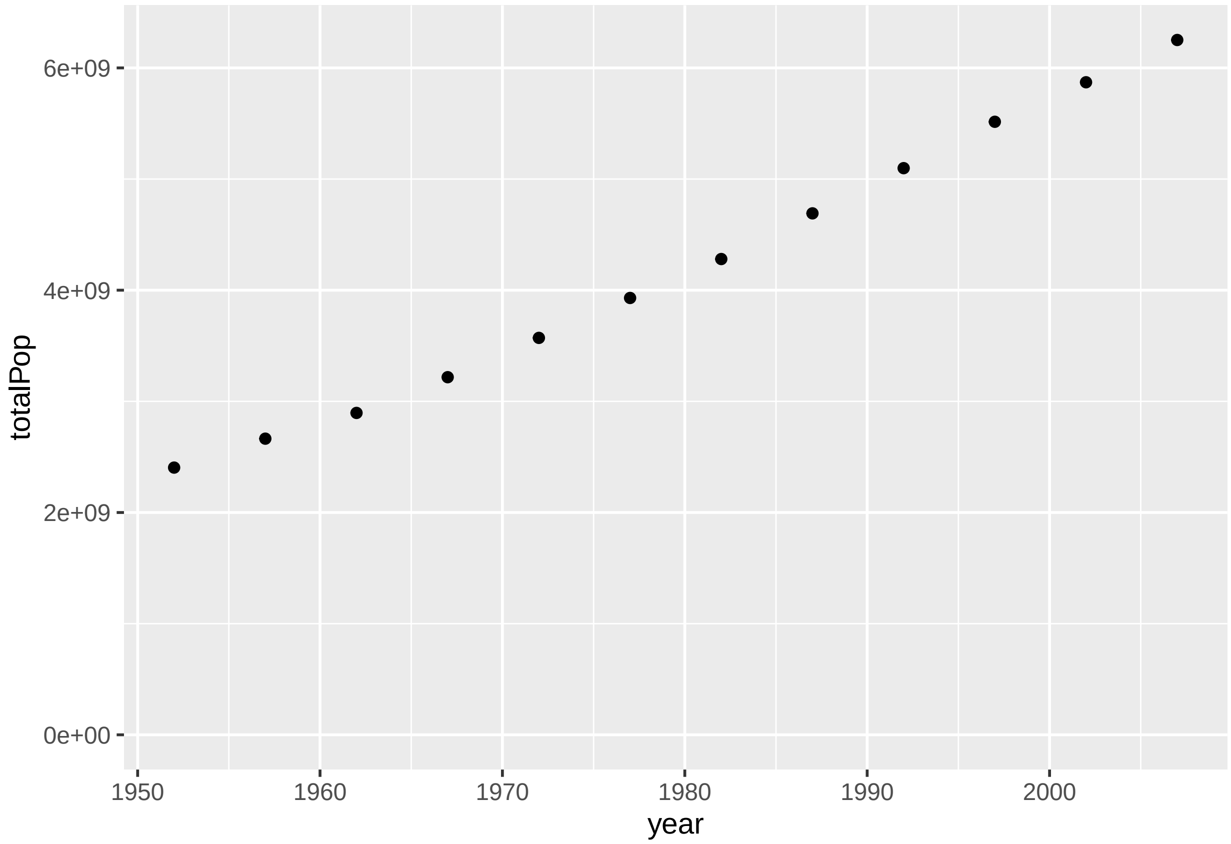
```
## # A tibble: 6 x 4
## # Groups:   continent [1]
##   continent  year medianLifeExp maxGdpPercap
##   <fct>     <int>         <dbl>        <dbl>
## 1 Africa     1952          38.6        4725.
## 2 Africa     1957          40.6        5487.
## 3 Africa     1962          42.6        6757.
## 4 Africa     1967          44.7       18773.
## 5 Africa     1972          47.0       21011.
## 6 Africa     1977          49.3       21951.
```

Visualizing summarized data

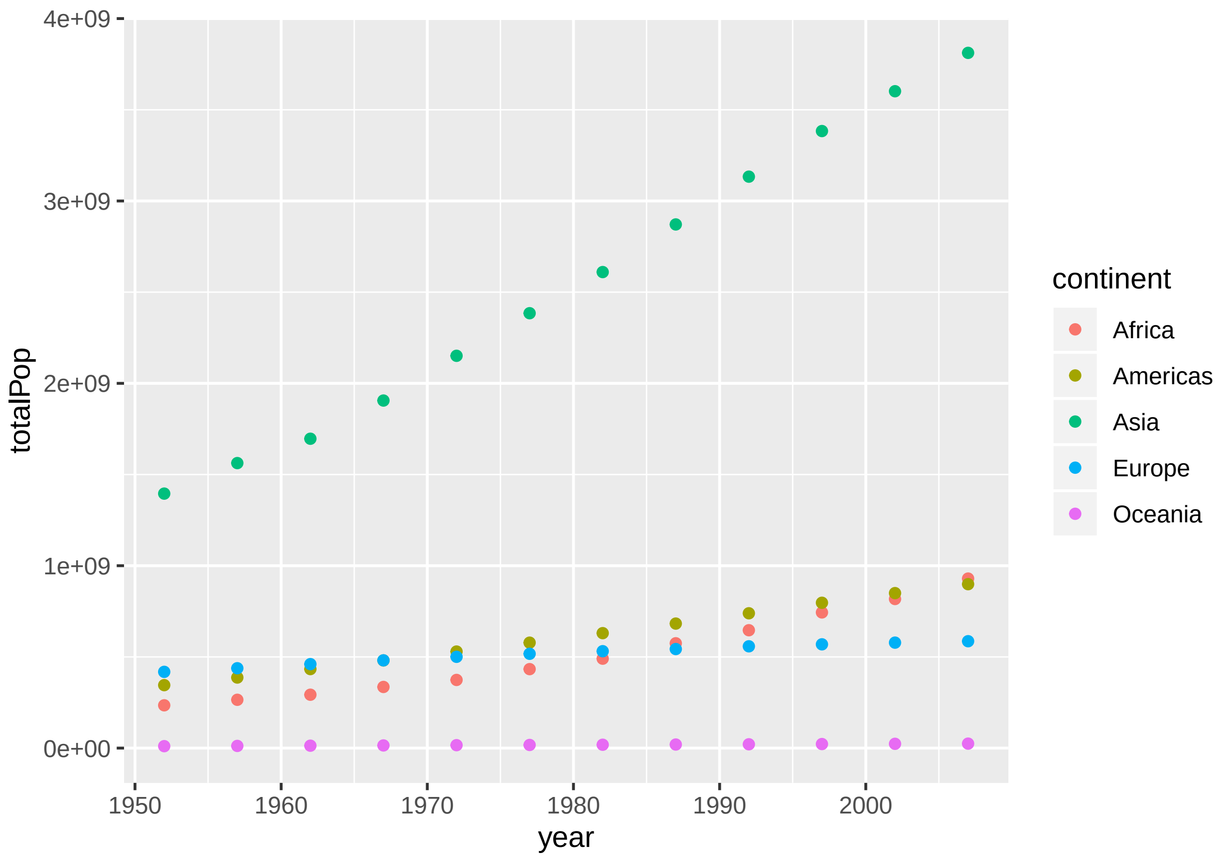
```
# Summarizing by year
by_year <- gapminder %>%
  group_by(year) %>%
  summarize(totalPop = sum(as.numeric(pop)),meanLifeExp = mean(lifeExp))

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)
```



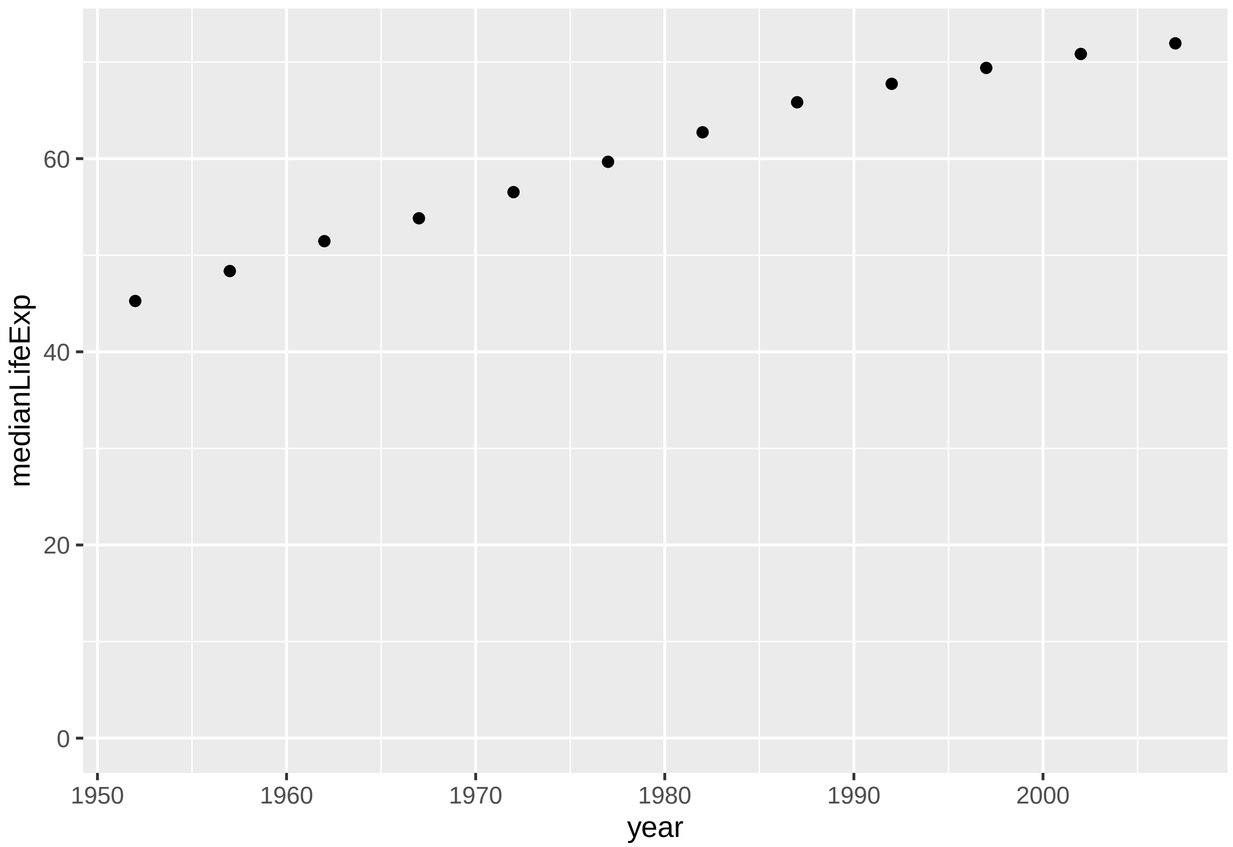
```
# Summarizing by year and continent
by_year_continent <- gapminder %>%
  group_by(year, continent) %>%
  summarize(totalPop = sum(as.numeric(pop)), meanLifeExp = mean(lifeExp))

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



Practice

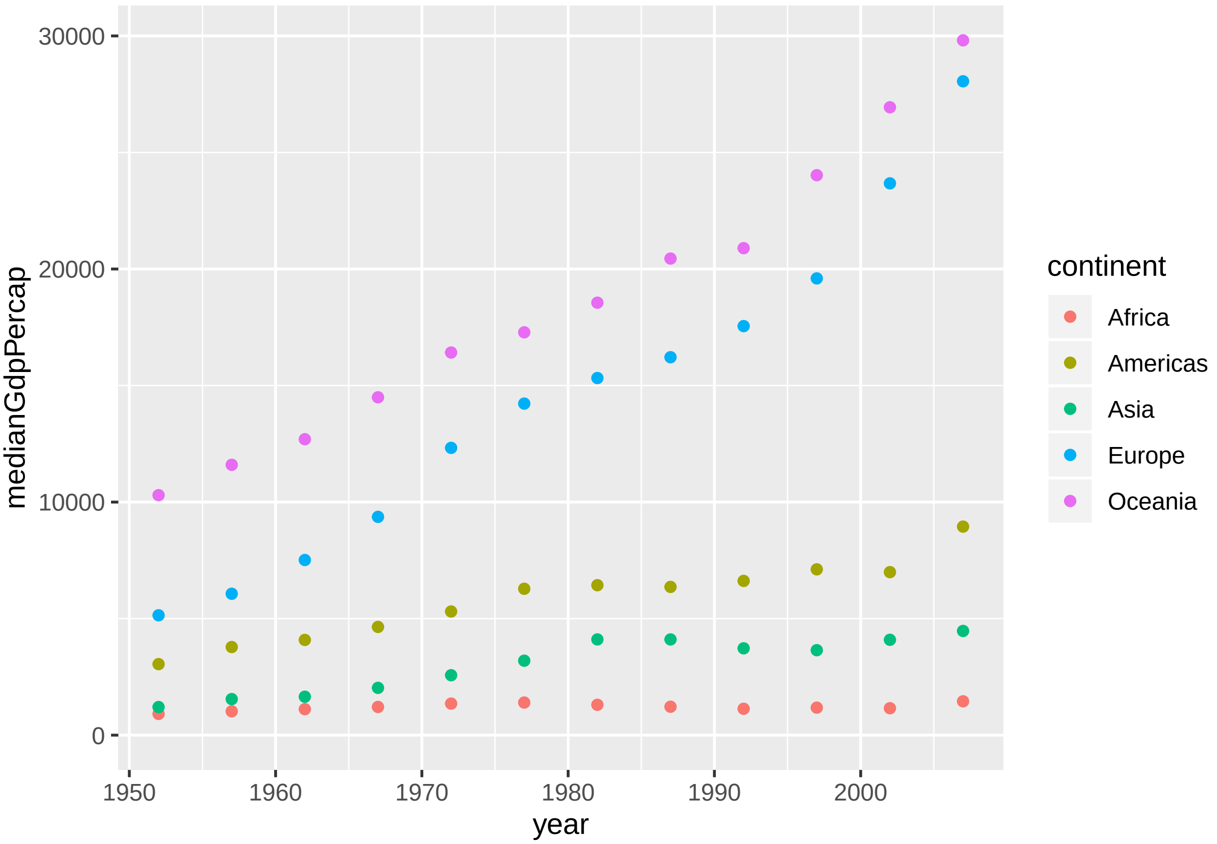
Visualizing median life expectancy over time

```
by_year <- gapminder %>%
  group_by(year) %>%
  summarize(medianLifeExp = median(lifeExp),
            maxGdpPercap = max(gdpPercap))

# Create a scatter plot showing the change in medianLifeExp over time
ggplot(by_year,aes(x = year, y = medianLifeExp)) +
  geom_point() +
  expand_limits(y = 0)
```



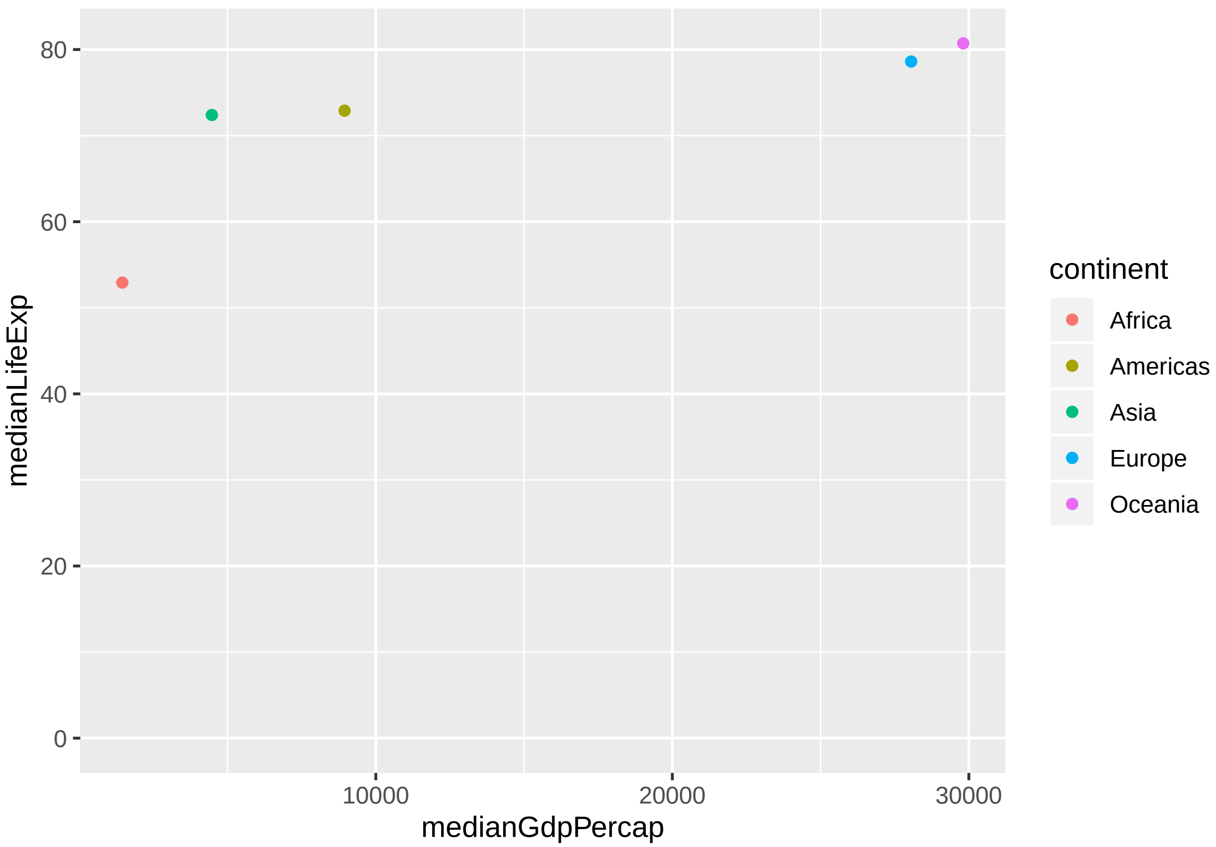
Visualizing median GDP per capita per continent over time

```
# Summarize medianGdpPercap within each continent within each year: by_year_continent
by_year_continent <- gapminder %>% group_by(continent,year) %>% summarize(medianGdpPercap = median(gdpPe

# Plot the change in medianGdpPercap in each continent over time
ggplot(by_year_continent,aes(x = year , y = medianGdpPercap, color = continent))+
  geom_point()+
  expand_limits(y = 0)
```



Comparing median life expectancy and median GDP per continent in 2007

```
# Summarize the median GDP and median life expectancy per continent in 2007
by_continent_2007 <- gapminder %>% filter(year == 2007) %>% group_by(continent) %>% summarize(medianLif

# Use a scatter plot to compare the median GDP and median life expectancy
ggplot(by_continent_2007, aes(x = medianGdpPercap, y = medianLifeExp, color = continent)) + geom_point(
```


80
60
40
20
0
medianLifeExp
10000
20000
30000
medianGdpPercap
continent
Africa
Americas
Asia
Europe
Oceania