hw02_dplyr_exploration

Haohui Zhong

2019/9/23

Outline

When exploring datasets, `dplyr` commands can be a very useful tool. Here, we would be focusing on the `gapminder` dataset. In order to do so, we need to load the `gapminder` package as well as the `tidyverse` package.

```
## -- Attaching packages ----------------------- tidyverse 1.2.1 --
## v ggplot2 3.2.0     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     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()
```

Exercise 1: Basic `dplyr`

1.1 `filter()`

Requirement: Use `filter()` to subset the `gapminder` data to three countries of our choice in the 1970s. Here I choose **China**, **Chad** and **Chile**.

```
gapminder %>%
  filter(year > 1969 & year < 1980) %>%
  filter(country == "China" | country == "Chad" | country == "Chile") %>%
  knitr::kable()
```

country	continent	year	lifeExp	pop	gdpPercap
Chad	Africa	1972	45.56900	3899068	1104.1040
Chad	Africa	1977	47.38300	4388260	1133.9850
Chile	Americas	1972	63.44100	9717524	5494.0244
Chile	Americas	1977	67.05200	10599793	4756.7638
China	Asia	1972	63.11888	862030000	676.9001
China	Asia	1977	63.96736	943455000	741.2375

1.2 `select()`

Requirement: Use the pipe operator '%>%' to select '"country"' and '"gdpPercap"' from our filtered dataset in *1.1*.

```
gapminder %>%
  filter(year > 1969 & year < 1980) %>%
  filter(country == "China" | country == "Chad" | country == "Chile") %>%
  select(country,gdpPercap) %>%
  knitr::kable()
```

country	gdpPercap
Chad	1104.1040
Chad	1133.9850
Chile	5494.0244
Chile	4756.7638
China	676.9001
China	741.2375

1.3 `mutate()`

Requirement: Filter `gapminder` to all entries that have experienced a drop in life expectancy and include a new variable that's the increase in life expectancy in the tibble.

```
gapminder %>%
  group_by(country) %>%
  arrange(year) %>%
  mutate(inc_lifeExp = lifeExp - lag(lifeExp)) %>%
  filter(inc_lifeExp < 0) %>%
  DT::datatable()
```

1.4 `max()`

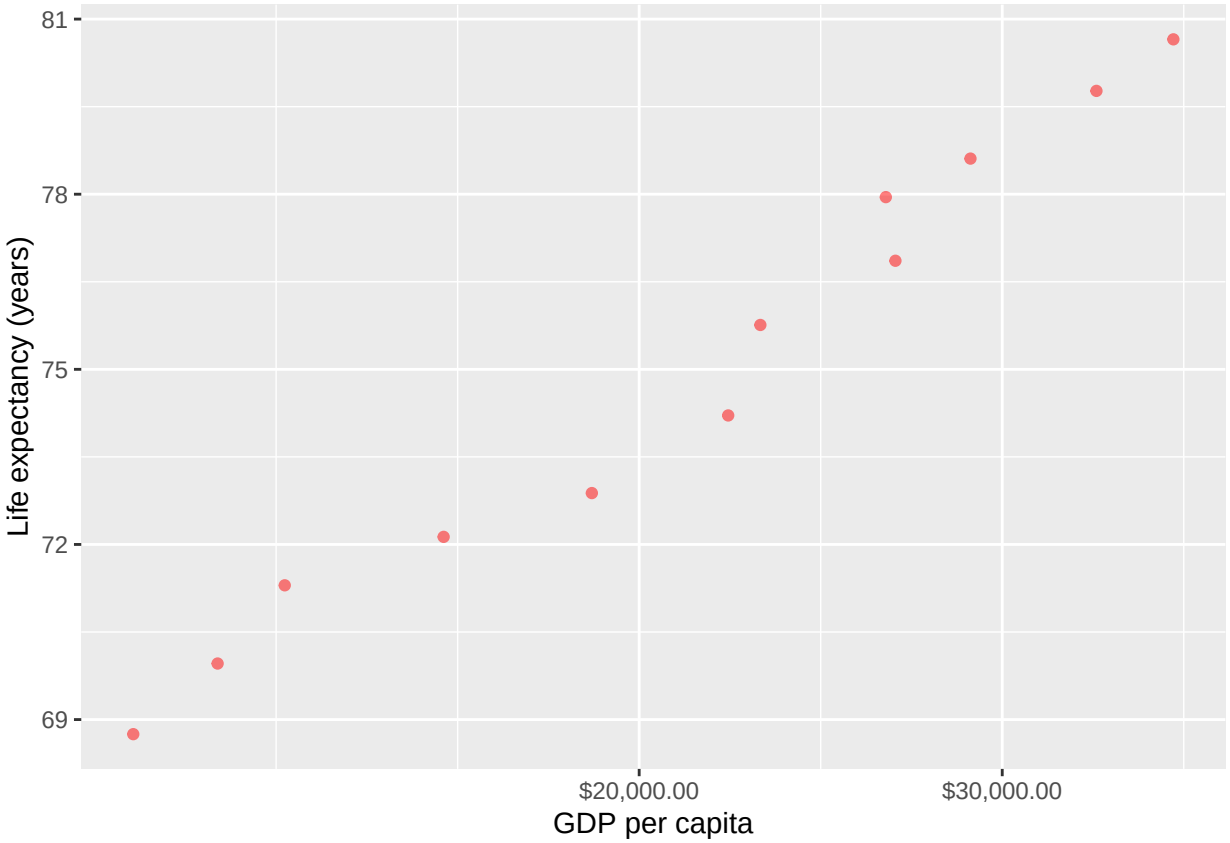
Requirement: Filter `gapminder` in order to show the max GDP per capita experienced by each country.

```
gapminder %>%
  group_by(country) %>%
  mutate(max_gdpPercap = max(gdpPercap)) %>%
  filter(gdpPercap == max_gdpPercap) %>%
  select(country, year, max_gdpPercap) %>%
  DT::datatable()
```

1.5 `ggplot()`

Requirement: Produce a scatterplot of Canada's life expectancy vs. GDP per capita using ggplot2, **without defining a new variable**. That is, after filtering the `gapminder` data set, pipe it directly into the `ggplot()` function. Ensure GDP per capita is on a log scale.

```
gapminder %>%
  filter(country == 'Canada') %>%
  ggplot(aes(gdpPercap, lifeExp)) +
  geom_point(alpha = 0.5, color = 'red') +
  scale_x_log10("GDP per capita", labels =  scales::dollar_format()) +
  ylab("Life expectancy (years)")
```



Exercise 2: Explore variables with `dplyr`

Requirements: Pick one categorical variable and one quantitative variable to explore. Answer the following questions using `dplyr`:

- What are possible values (or range) of each variable?
- What values are typical? What's the spread? What's the distribution?

Here I choose the *categorical* variable `'continent'` and the *quantitative* variable `'lifeExp'`.

2.1 categorical variable `'continent'`

To investigate what possible values of such categorical variable are, we can study the variable as well as its frequency.

```
levels(gapminder$continent)
```

```
## [1] "Africa"   "Americas" "Asia"     "Europe"   "Oceania"
```

```
gapminder %>%
  count(continent)
```

```
## # A tibble: 5 x 2
##   continent     n
##   <fct>     <int>
## 1 Africa      624
## 2 Americas    300
## 3 Asia        396
```

```
## 4 Europe      360
## 5 Oceania      24
```

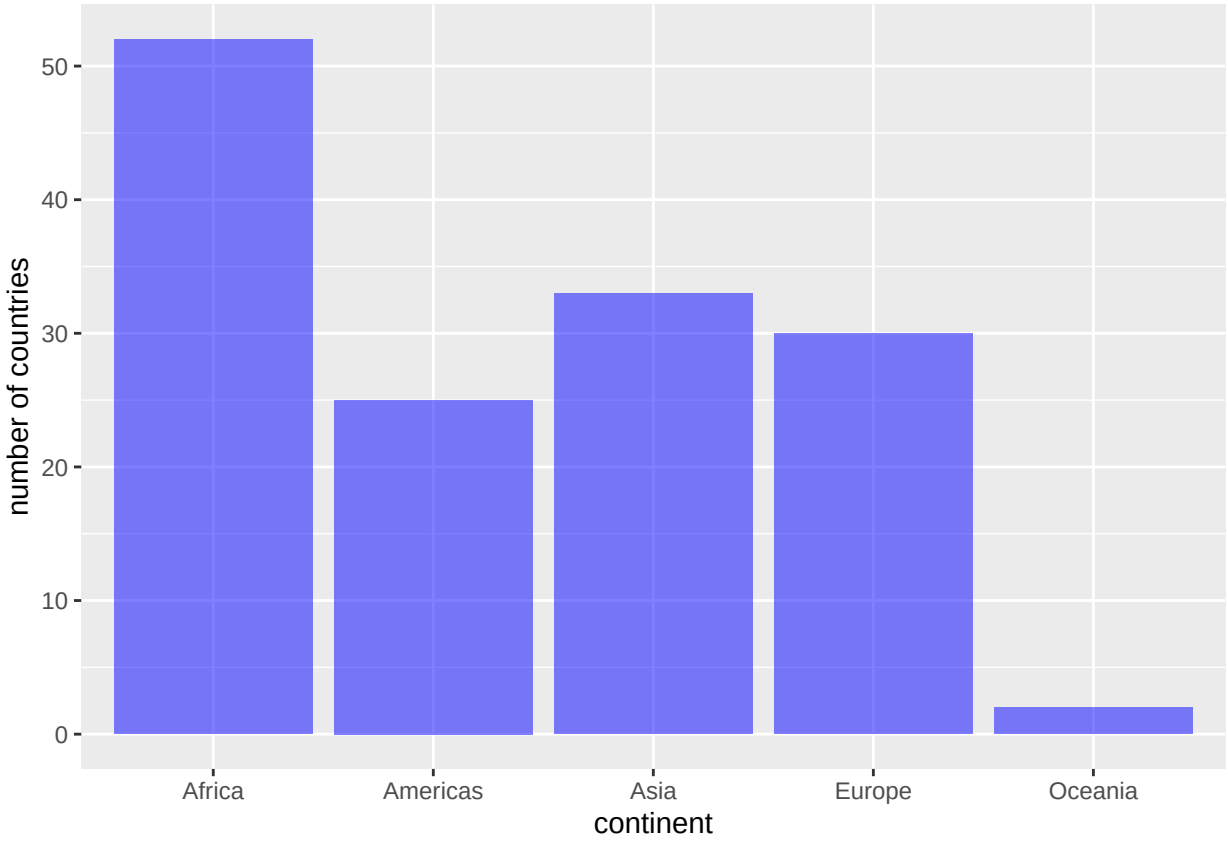
The above chunk shows numbers of observations for each continent. But if we check back the `gapminder` dataset, it is noticeable that each country of the corresponding continent contributes multiple records. Yet we are not sure whether numbers of countries within each continent have changed over time.

```
gapminder %>%
  select(continent,year) %>%
  group_by(year) %>%
  count(continent)
```

```
## # A tibble: 60 x 3
## # Groups:   year [12]
##     year continent     n
##    <int> <fct>     <int>
##  1  1952 Africa       52
##  2  1952 Americas     25
##  3  1952 Asia         33
##  4  1952 Europe       30
##  5  1952 Oceania       2
##  6  1957 Africa       52
##  7  1957 Americas     25
##  8  1957 Asia         33
##  9  1957 Europe       30
## 10  1957 Oceania       2
## # ... with 50 more rows
```

This confirms that numbers of countries in each continent remain the same. Therefore, we can get rid of redundant repeats of each country. Here, we can learn that `continent` Africa contains 52 countries, Americas contains 25, Asia contains 33, Europe contains 30 and Oceania 2. To better view such result, we can visualize it.

```
gapminder %>%
  filter(year == 1952) %>%
  ggplot(aes(continent)) +
  geom_bar(fill = "blue", alpha = 0.5) +
  labs(x="continent", y="number of countries")
```



2.2 quantitative variable 'lifeExp'

We can use `summary()` to have a very brief idea of the variable `'lifeExp'`.

```
summary(gapminder$lifeExp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   23.60   48.20   60.71   59.47   70.85   82.60
```

We can also learn information about the standard deviation of these values.

```
sd(gapminder$lifeExp)
```

```
## [1] 12.91711
```

The above information is very general, without distinguishing observations among years, countries and continents. Life expectancy might change over time and vary among countries. We can have a closer look at these values by grouping them into different subsets.

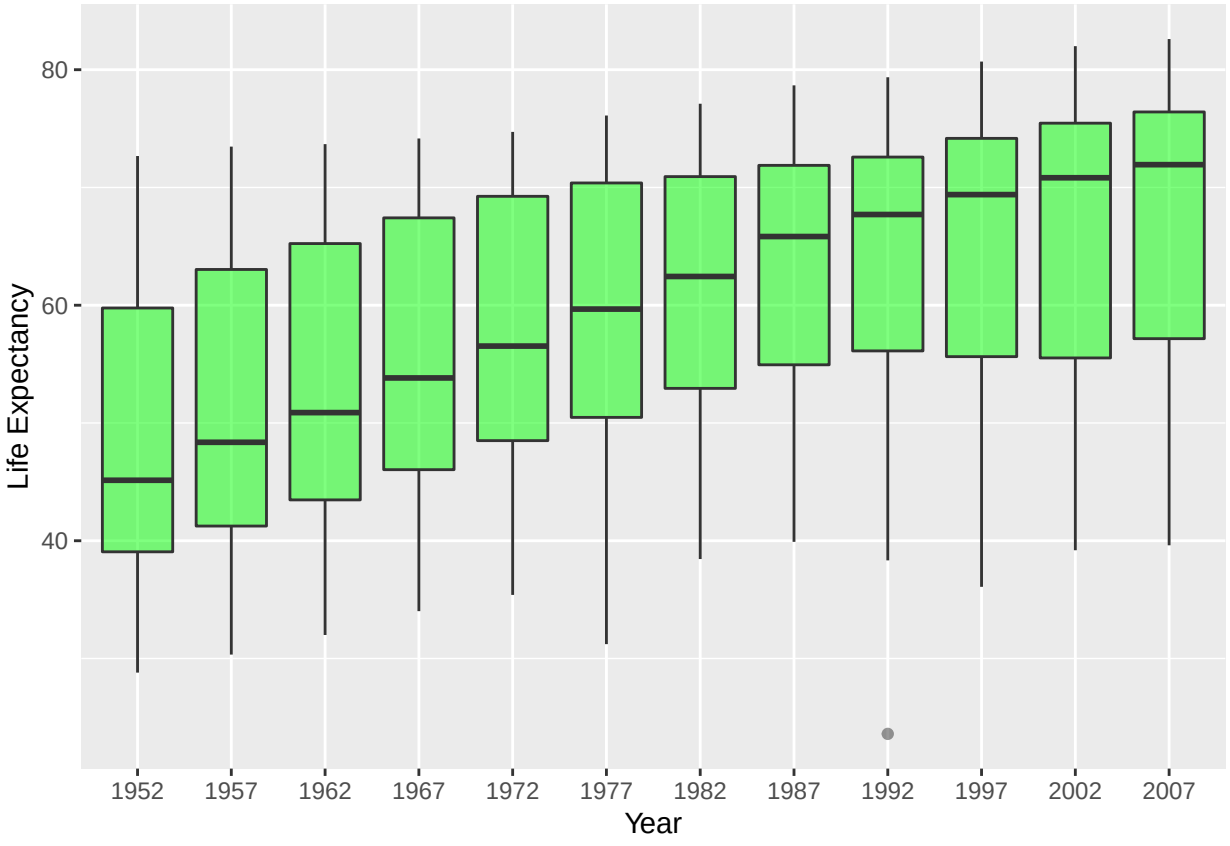
```
gapminder %>%
  select(year, lifeExp) %>%
  group_by(year) %>%
  arrange(year) %>%
  summarize(min_lifeExp = min(lifeExp), max_lifeExp = max(lifeExp), mean_lifeExp = mean(lifeExp),median_
  knitr::kable()
```

year	min_lifeExp	max_lifeExp	mean_lifeExp	median_lifeExp	sd_lifeExp
1952	28.801	72.670	49.05762	45.1355	12.22596

year	min_lifeExp	max_lifeExp	mean_lifeExp	median_lifeExp	sd_lifeExp
1957	30.332	73.470	51.50740	48.3605	12.23129
1962	31.997	73.680	53.60925	50.8810	12.09724
1967	34.020	74.160	55.67829	53.8250	11.71886
1972	35.400	74.720	57.64739	56.5300	11.38195
1977	31.220	76.110	59.57016	59.6720	11.22723
1982	38.445	77.110	61.53320	62.4415	10.77062
1987	39.906	78.670	63.21261	65.8340	10.55629
1992	23.599	79.360	64.16034	67.7030	11.22738
1997	36.087	80.690	65.01468	69.3940	11.55944
2002	39.193	82.000	65.69492	70.8255	12.27982
2007	39.613	82.603	67.00742	71.9355	12.07302

Visual aids are always helpful to understant trends and diviations of values.

```
gapminder %>%
  select(year, lifeExp) %>%
  mutate(year= factor(year)) %>%
  ggplot(aes(year, lifeExp)) +
  geom_boxplot(alpha = 0.5, fill = "green") +
  xlab("Year") +
  ylab("Life Expectancy")
```



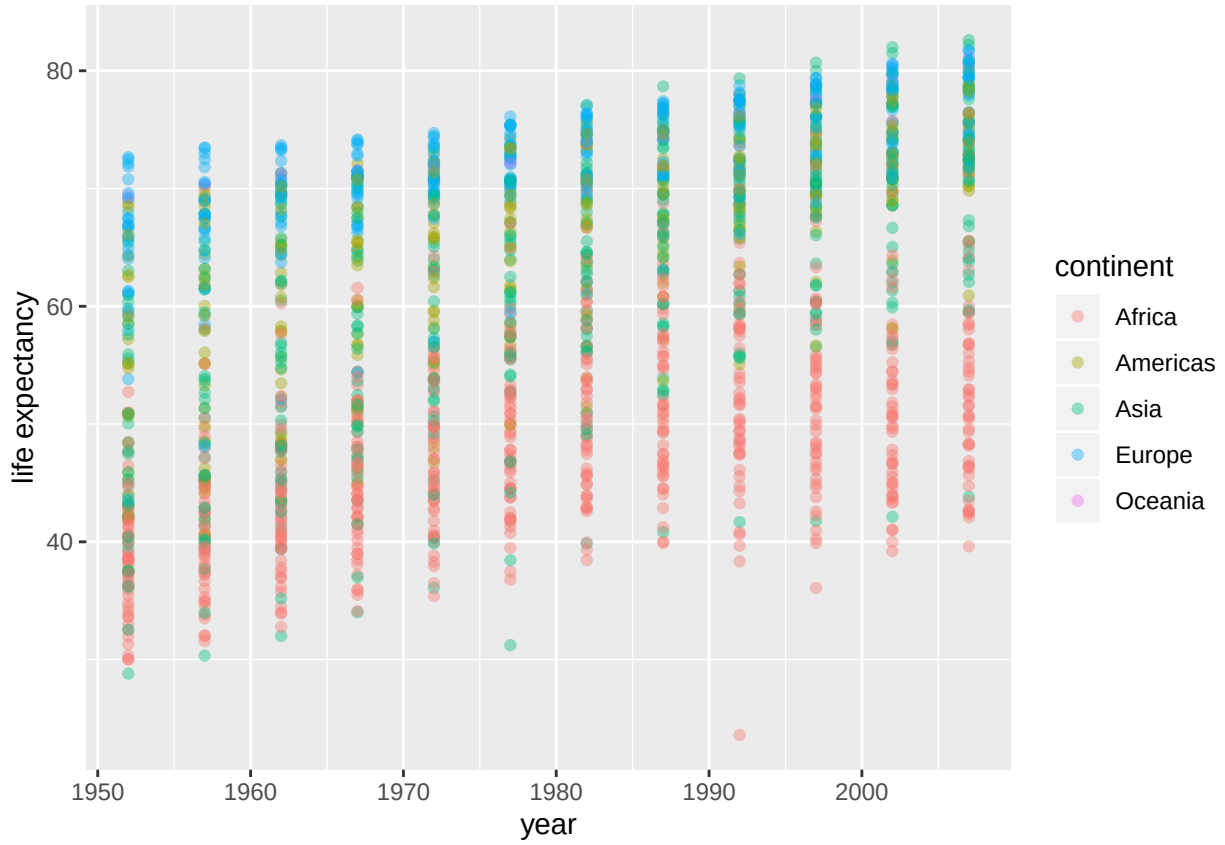
Exercise 3: Explore various plot types

Requirements: Make the following plots that could be used for analyses: - A scatterplot of two quantitative variables. - One other plot besides a scatterplot.

3.1 Scatterplot

In *Exercise 2*, our exploration reveals that life expectancy changes over time. Whether life expectancy differs among continents is unclear. By studying relationships among these three variables, we might learn some interesting facts and see how things develop.

```
gapminder %>%
  group_by(continent) %>%
  ggplot(aes(color = continent,x = year, y = lifeExp))+
  geom_point(alpha = 0.4) +
  xlab("year")+
  ylab("life expectancy")
```



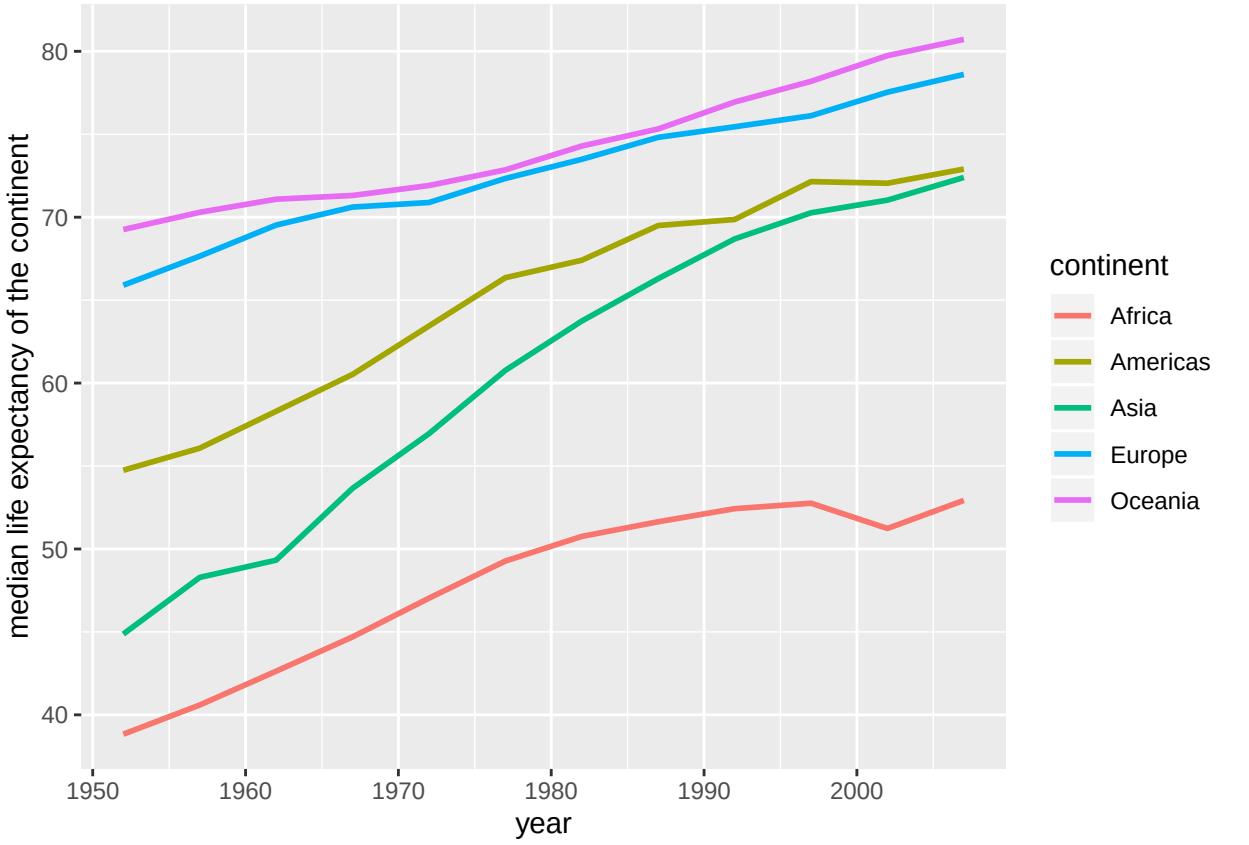
Here we can see that throughout the years Europe enjoys an extremely long life expectancy, followed by Americas, while life expectancy in Asia remains the shortest with gradual increases.

3.2 Other types of plots

Instead of visualizing observations from every country in all continents, we can narrow down our choices. For example, we study how median life expectancy within each continent change over the years.

```
gapminder %>%
  group_by(continent,year) %>%
```

```
  mutate(md_lifeExp = median(lifeExp)) %>%
  ggplot(aes(x = year, y = md_lifeExp, color = continent))+
  geom_line(size=1)+
  xlab("year")+
  ylab("median life expectancy of the continent")
```



Optional Exercise: Recycling

The analyst's propose was to obtain data of Rwanda and Afghanistan, and yet outputs of the following chunk did not fulfill the requirement. This is mainly due to the recycling of the vector `c("Rwanda", "Afghanistan")`, which results in observations of the two countries taking turns to be obtained year by year.

```
filter(gapminder, country == c("Rwanda", "Afghanistan"))
```

```
## # A tibble: 12 x 6
##    country     continent  year lifeExp      pop gdpPercap
##    <fct>       <fct>     <int>   <dbl>    <int>     <dbl>
##  1 Afghanistan Asia       1957    30.3  9240934      821.
##  2 Afghanistan Asia       1967    34.0 11537966      836.
##  3 Afghanistan Asia       1977    38.4 14880372      786.
##  4 Afghanistan Asia       1987    40.8 13867957      852.
##  5 Afghanistan Asia       1997    41.8 22227415      635.
##  6 Afghanistan Asia       2007    43.8 31889923      975.
##  7 Rwanda      Africa     1952    40    2534927      493.
##  8 Rwanda      Africa     1962    43    3051242      597.
##  9 Rwanda      Africa     1972    44.6  3992121      591.
```

```
## 10 Rwanda      Africa     1982    46.2  5507565      882.
## 11 Rwanda      Africa     1992    23.6  7290203      737.
## 12 Rwanda      Africa     2002    43.4  7852401      786.
```

To obtain the values properly, all we need to do is just to slightly modify the code.

```
filter(gapminder, country == "Rwanda" | country == "Afghanistan") %>%
DT::datatable()
```