# Introduction to dplyr

`select()`

`mutate()`

on columns

`arrange()`

`filter()`

| | Variable1 | Variable2 | Variable3 | Variable4 | Variable5 |
|---|---|---|---|---|---|
| case1 | | | | | |
| case2 | | | | | |
| case3 | | | | | |
| case4 | | | | | |

`summarize()`

```
arrange(gapm, Year, Country)

arrange(gapm, desc(lifeExp), Year, continent)
```

| | Country | Year | pop | continent | lifeExp | gdpPercap |
|---|---|---|---|---|---|---|
| case1 | | | | | | |
| case2 | | | | | | |
| case3 | | | | | | |
| case4 | | | | | | |

Changes order

Ordered by first column name, then ties resolved with second etc.

Do not quote column names!

For decreasing order, use `desc()`

```
filter(gapm, Year >1993 , Country==“UK”)

filter(gapm, gdpPercap*pop >10^7 , Year ==1993 )
```

| | Country | Year | pop | continent | lifeExp | gdpPercap |
|---|---|---|---|---|---|---|
| case1 | | | | | | |
| case2 | | | | | | |
| case3 | | | | | | |
| case4 | | | | | | |

Do not quote column names!

```
select(gapm, 1)
select(gapm, Country, continent)
select(gapm, Country: pop)
select(gapm, -(continent, lifeExp))
```

| | Country | Year | pop | continent | lifeExp | gdpPercap |
|---|---|---|---|---|---|---|
| case1 | | | | | | |
| case2 | | | | | | |
| case3 | | | | | | |
| case4 | | | | | | |

Select helpers

- starts_with(): starts with a prefix
- ends_with(): ends with a prefix
- contains(): contains a literal string
- one_of(): variables in character vector.

```
mutate(gapm, totalGDP= gdpPercap * pop)
mutate(gapm, totalGDP= gdpPercap * pop,
         totalGDPpounds= totalGDP *0.76)
```

| | Country | Year | pop | continent | lifeExp | gdpPercap |
|---|---|---|---|---|---|---|
| case1 | | | | | | |
| case2 | | | | | | |
| case3 | | | | | | |
| case4 | | | | | | |

transmute()

| | Country | Year | pop | continent | lifeExp | gdpPercap |
|---|---|---|---|---|---|---|
| case1 | | | | | | |
| case2 | | | | | | |
| case3 | | | | | | |
| case4 | | | | | | |

```
summarise(gapm, min_lifeExp = min(lifeExp, na.rm = TRUE) )
```

A function which would take a vector of values and return 1 value

min(), max(), mean(), sum(), sd(), median()
n(), n_distinct(), first(x), last(x) , nth(x, n)

```
By_Country <- group_by(gapm, Country )
Summarise(By_Country, max(lifeExp))
```

```
group_by(gapm, Year, continent )
```

| | Country | Year | pop | continent | lifeExp | gdpPercap |
|---|---|---|---|---|---|---|
| case1 | | | | | | |
| case2 | | | | | | |
| case3 | | | | | | |
| case4 | | | | | | |

Operations on groups of rows

# %>%

Pass the output directly to the next function

```
A <- select(df, …)
B <- mutate(A, …)
C <- filter(B, …)
```

```
filter(mutate( select(df, …) , …) , …)
```

```
select(df, …) %>% mutate(…) %>% filter(…)
```