Introduction to `dplyr` - Solutions

Exercise A - (3 min)

1. Install the `dplyr` and `gapminder` packages.
2. Load the `dplyr` and `gapminder` packages.
3. Read the help file for `gapminder` and answer the following:
    a. How many rows and columns does `gapminder` contain?
    b. What information is in each row and column?
    c. What is the source of the data?

Solution

1. Run the following:

```
install.packages('dplyr')
install.packages('gapminder')
```

2. Run the following:

```
library(dplyr)
library(gapminder)
```

3. Consult `?gapminder`.

Exercise B - (5 minutes)

1. What is the difference between `x = 3` and `x == 3` in R?
2. On the previous slide, I put quotes around `United States` but not around `year`. Why?
3. Use `filter` to choose the subset of `gapminder` for 2002. What happens if you replace 2002 with 2005?
4. Store data for all Asian countries in a tibble called `gapminder_asia`, then display this tibble.
5. Which country had the higher life expectancy in 1977: Ireland or Brazil? Which had the higher GDP per capita?

Solution

1. The first *assigns* the value `3` to `x`; the second *tests* whether or not `x` is equal to `3` and returns `TRUE` or `FALSE`
2. Because `year` contains numeric data while `country` contains character data.
3. If you go back to the help file for `gapminder` you'll see that it only contains data for every fifth year. The year 2005 isn't in our so `dplyr` displayes an empty tibble in the second case:

```
gapminder |>
  filter(year == 2002)
```

```
# A tibble: 142 × 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.
# ℹ 132 more rows
```

```
gapminder |>
  filter(year == 2005)
```

```
# A tibble: 0 × 6
# ℹ 6 variables: country <fct>, continent <fct>, year <int>, lifeExp <dbl>,
#   pop <int>, gdpPercap <dbl>
```

4. Run the following:

```
gapminder_asia <- gapminder |>
  filter(continent == 'Asia')
gapminder_asia
```

```
# A tibble: 396 × 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.
# ℹ 386 more rows
```

5. Ireland had the higher value of both:

```
gapminder |>
  filter(year == 1977, country %in% c('Ireland', 'Brazil'))
```

```
# A tibble: 2 × 6
  country continent  year lifeExp       pop gdpPercap
  <fct>   <fct>     <int>   <dbl>     <int>     <dbl>
1 Brazil  Americas   1977    61.5 114313951     6660.
2 Ireland Europe     1977    72.0   3271900    11151.
```

Exercise C - (2 min)

1. What is the lowest life expectancy in `gapminder`? Which country and year does it correspond to?
2. What is the highest life expectancy in `gapminder`? Which country and year does it correspond to?

Solution

1. The lowest life expectancy was in Rwanda in 1992:

```
gapminder |>
  arrange(lifeExp)
```

```
# A tibble: 1,704 × 6
   country      continent  year lifeExp     pop gdpPercap
   <fct>        <fct>     <int>   <dbl>   <int>     <dbl>
 1 Rwanda       Africa     1992    23.6 7290203      737.
 2 Afghanistan  Asia       1952    28.8 8425333      779.
 3 Gambia       Africa     1952    30    284320      485.
 4 Angola       Africa     1952    30.0 4232095     3521.
 5 Sierra Leone Africa     1952    30.3 2143249      880.
 6 Afghanistan  Asia       1957    30.3 9240934      821.
 7 Cambodia     Asia       1977    31.2 6978607      525.
 8 Mozambique   Africa     1952    31.3 6446316      469.
 9 Sierra Leone Africa     1957    31.6 2295678     1004.
10 Burkina Faso Africa     1952    32.0 4469979      543.
# ℹ 1,694 more rows
```

2. The highest life expectancy was in Japan in 2007:

```
gapminder |>
  arrange(desc(lifeExp))
```

```
# A tibble: 1,704 × 6
   country          continent  year lifeExp       pop gdpPercap
   <fct>            <fct>     <int>   <dbl>     <int>     <dbl>
 1 Japan            Asia       2007    82.6 127467972    31656.
 2 Hong Kong, China Asia       2007    82.2   6980412    39725.
 3 Japan            Asia       2002    82   127065841    28605.
 4 Iceland          Europe     2007    81.8    301931    36181.
 5 Switzerland      Europe     2007    81.7   7554661    37506.
 6 Hong Kong, China Asia       2002    81.5   6762476    30209.
 7 Australia        Oceania    2007    81.2  20434176    34435.
 8 Spain            Europe     2007    80.9  40448191    28821.
 9 Sweden           Europe     2007    80.9   9031088    33860.
10 Israel           Asia       2007    80.7   6426679    25523.
# ℹ 1,694 more rows
```

Exercise D - (2 min)

1. Select only the columns `year`, `lifeExp`, and `country` in `gapminder`.
2. Select all the columns *except* `year`, `lifeExp` and `country` in `gapminder`.

Solution

```
# Part 1
gapminder |>
  select(year, lifeExp, country)
```

```
# A tibble: 1,704 × 3
    year lifeExp country
   <int>   <dbl> <fct>
 1  1952    28.8 Afghanistan
 2  1957    30.3 Afghanistan
 3  1962    32.0 Afghanistan
 4  1967    34.0 Afghanistan
 5  1972    36.1 Afghanistan
 6  1977    38.4 Afghanistan
 7  1982    39.9 Afghanistan
 8  1987    40.8 Afghanistan
 9  1992    41.7 Afghanistan
10  1997    41.8 Afghanistan
# ℹ 1,694 more rows
```

```
# Part 2
gapminder |>
  select(-year, -lifeExp, -country)
```

```
# A tibble: 1,704 × 3
   continent      pop gdpPercap
   <fct>        <int>     <dbl>
 1 Asia       8425333      779.
 2 Asia       9240934      821.
 3 Asia      10267083      853.
 4 Asia      11537966      836.
 5 Asia      13079460      740.
 6 Asia      14880372      786.
 7 Asia      12881816      978.
 8 Asia      13867957      852.
 9 Asia      16317921      649.
10 Asia      22227415      635.
# ℹ 1,694 more rows
```

Exercise E - (2 min)

1. Compute the median life expectancy in 1977.
2. Repeat 1 but restrict the calculation to Asian countries.

Solution

```
# Part 1
gapminder |>
  filter(year == 1977) |>
  summarize(median(lifeExp))
```

```
# A tibble: 1 × 1
  `median(lifeExp)`
              <dbl>
1              59.7
```

```
# Part 2
gapminder |>
  filter(year == 1977, continent == 'Asia') |>
  summarize(median(lifeExp))
```

```
# A tibble: 1 × 1
  `median(lifeExp)`
              <dbl>
1              60.8
```

Exercise F - (2 min)

1. Calculate median GDP/capita in each continent in 1977.
2. Why doesn't this work as expected? How can you fix it?

```
gapminder |>
  summarize(meanLifeExp = mean(lifeExp)) |>
  group_by(year)
```

Solution

1. Run the following:

```
gapminder |>
  group_by(continent) |>
  summarize(median(lifeExp))
```

```
# A tibble: 5 × 2
  continent `median(lifeExp)`
  <fct>                 <dbl>
1 Africa                 47.8
2 Americas               67.0
3 Asia                   61.8
```

```
 8 Afghanistan Asia       1987    40.8 13867957      852.          490.
 9 Afghanistan Asia       1992    41.7 16317921      649.          500.
10 Afghanistan Asia       1997    41.8 22227415      635.          501.
# ℹ 1,694 more rows
```

Exercise H - (2 min)

1. Use `|>` to calculate the sample variance of `c(4, 1, 5, NA, 3)`, excluding any missing values.
2. Repeat the preceding using *both* `|>` and `_`.
3. Sort `gapminder` in descending order by `lifeExp` *without* using `|>` or `_`.

Solution

```
# Part 1
c(4, 1, 5, NA, 3) |>
  var(na.rm = TRUE)
```

```
[1] 2.916667
```

```
# Part 2
TRUE |>
  var(c(4, 1, 5, NA, 3), na.rm = _)
```

```
[1] 2.916667
```

```
# Part 3
arrange(gapminder, desc(lifeExp))
```

```
# A tibble: 1,704 × 6
   country          continent  year lifeExp       pop gdpPercap
   <fct>            <fct>     <int>   <dbl>     <int>     <dbl>
 1 Japan            Asia       2007    82.6 127467972    31656.
 2 Hong Kong, China Asia       2007    82.2   6980412    39725.
 3 Japan            Asia       2002    82   127065841    28605.
 4 Iceland          Europe     2007    81.8    301931    36181.
 5 Switzerland      Europe     2007    81.7   7554661    37506.
 6 Hong Kong, China Asia       2002    81.5   6762476    30209.
 7 Australia        Oceania    2007    81.2  20434176    34435.
 8 Spain            Europe     2007    80.9  40448191    28821.
 9 Sweden           Europe     2007    80.9   9031088    33860.
10 Israel           Asia       2007    80.7   6426679    25523.
# ℹ 1,694 more rows
```

Exercise I - (5 min)

Write a single pipeline that calculates the mean and standard deviation of GDP/capita by continent and year for all years *after* 1997, and sorts the results in ascending order by the standard deviation.

```
4 Europe                 72.2
5 Oceania                73.7
```

2. Here the problem is that `group_by()` comes *after* `summarize()`, but once we've summarized by computing the mean life expectancy, we've already "collapsed" all the years. The code works as expected if we reverse the order:

```
gapminder |>
  group_by(year) |>
  summarize(meanLifeExp = mean(lifeExp))
```

```
# A tibble: 12 × 2
    year meanLifeExp
   <int>       <dbl>
 1  1952        49.1
 2  1957        51.5
 3  1962        53.6
 4  1967        55.7
 5  1972        57.6
 6  1977        59.6
 7  1982        61.5
 8  1987        63.2
 9  1992        64.2
10  1997        65.0
11  2002        65.7
12  2007        67.0
```

Exercise G - (2min)

1. Why did I use `=` rather than `==` in the `mutate()` examples from the preceding two slides?
2. Convert life expectancy from years to *months*.

Solution

1. This is because we are carrying out an *assignment operation*. In contrast, `==` tests for equality, returning `TRUE` or `FALSE`.
2. Run the following:

```
gapminder |>
  mutate(lifeExpMonths = 12 * lifeExp)
```

```
# A tibble: 1,704 × 7
   country     continent  year lifeExp      pop gdpPercap lifeExpMonths
   <fct>       <fct>     <int>   <dbl>    <int>     <dbl>          <dbl>
 1 Afghanistan Asia       1952    28.8  8425333      779.           346.
 2 Afghanistan Asia       1957    30.3  9240934      821.           364.
 3 Afghanistan Asia       1962    32.0 10267083      853.           384.
 4 Afghanistan Asia       1967    34.0 11537966      836.           408.
 5 Afghanistan Asia       1972    36.1 13079460      740.           433.
 6 Afghanistan Asia       1977    38.4 14880372      786.           461.
 7 Afghanistan Asia       1982    39.9 12881816      978.           478.
```

Solution

```
gapminder |>
  filter(year > 1997) |>
  group_by(continent, year) |>
  summarize(mean_GDPc = mean(gdpPercap), sd_GDPc = sd(gdpPercap)) |>
  arrange(sd_GDPc)
```

```
# A tibble: 10 × 4
# Groups:   continent [5]
   continent  year mean_GDPc sd_GDPc
   <fct>     <int>     <dbl>   <dbl>
 1 Africa     2002     2599.   2973.
 2 Africa     2007     3089.   3618.
 3 Oceania    2002    26939.   5302.
 4 Oceania    2007    29810.   6541.
 5 Americas   2002     9288.   8896.
 6 Americas   2007    11003.   9713.
 7 Asia       2002    10174.  11151.
 8 Europe     2002    21712.  11197.
 9 Europe     2007    25054.  11800.
10 Asia       2007    12473.  14155.
```