R Practice 4 Answers

Ryan Safner

11/15/2018

Intro to dplyr syntax

1. Load the gapminder and tidyverse packages.

```
suppressPackageStartupMessages(library("tidyverse"))
library("gapminder")
```

select()

1. Make a data frame containing the columns year, lifeExp, country from the gapminder data, in that order.

```
select(gapminder, c(year, lifeExp, country))
```

```
## # A tibble: 1,704 x 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
## # ... with 1,694 more rows
```

```
# using the pipe

gapminder %>%
  select(c(year, lifeExp, country))
```

```
## # A tibble: 1,704 x 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
## # ... with 1,694 more rows
```

2. Select all variables, from `country` to `lifeExp`.

3. Select all variables, except `pop`.

```
select(gapminder, -pop)
```

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

```
# using the pipe

gapminder %>%
  select(-pop)
```

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

4. Rename `continent` to `cont`.

```
rename(gapminder, cont=continent)
```

```
## # A tibble: 1,704 x 6
##    country     cont   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
```

```
# using the pipe

gapminder %>%
  rename(cont=continent)
```

```
## # A tibble: 1,704 x 6
##    country     cont   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
```

arrange()

1. Order by year.

```
gapminder %>%
  arrange(year)
```

```
## # 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 Albania     Europe     1952    55.2  1282697     1601.
##  3 Algeria     Africa     1952    43.1  9279525     2449.
##  4 Angola      Africa     1952    30.0  4232095     3521.
##  5 Argentina   Americas   1952    62.5 17876956     5911.
##  6 Australia   Oceania    1952    69.1  8691212    10040.
##  7 Austria     Europe     1952    66.8  6927772     6137.
##  8 Bahrain     Asia       1952    50.9   120447     9867.
##  9 Bangladesh  Asia       1952    37.5 46886859      684.
## 10 Belgium     Europe     1952    68    8730405     8343.
## # ... with 1,694 more rows
```

2. Order by year, in descending order.

```r
gapminder %>%
  arrange(desc(year))
```

```
## # A tibble: 1,704 x 6
##    country     continent  year lifeExp       pop gdpPercap
##    <fct>       <fct>     <int>   <dbl>     <int>     <dbl>
##  1 Afghanistan Asia       2007    43.8  31889923      975.
##  2 Albania     Europe     2007    76.4   3600523     5937.
##  3 Algeria     Africa     2007    72.3  33333216     6223.
##  4 Angola      Africa     2007    42.7  12420476     4797.
##  5 Argentina   Americas   2007    75.3  40301927    12779.
##  6 Australia   Oceania    2007    81.2  20434176    34435.
##  7 Austria     Europe     2007    79.8   8199783    36126.
##  8 Bahrain     Asia       2007    75.6    708573    29796.
##  9 Bangladesh  Asia       2007    64.1 150448339     1391.
## 10 Belgium     Europe     2007    79.4  10392226    33693.
## # ... with 1,694 more rows
```

3. Order by year, then by life expectancy.

```r
gapminder %>%
  arrange(year, lifeExp)
```

```
## # 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 Gambia        Africa     1952    30    284320      485.
##  3 Angola        Africa     1952    30.0 4232095     3521.
##  4 Sierra Leone  Africa     1952    30.3 2143249      880.
##  5 Mozambique    Africa     1952    31.3 6446316      469.
##  6 Burkina Faso  Africa     1952    32.0 4469979      543.
##  7 Guinea-Bissau Africa     1952    32.5  580653      300.
##  8 Yemen, Rep.   Asia       1952    32.5 4963829      782.
##  9 Somalia       Africa     1952    33.0 2526994     1136.
## 10 Guinea        Africa     1952    33.6 2664249      510.
## # ... with 1,694 more rows
```

Piping, %>%

Note: think of %>% as the word "then"!

1. Subset your data to look only at year, gdpPercap, and country in the year 1997, for countries that have a gdpPercap greater than 20,000, and order them alphabetically.

```r
gapminder %>%
  select(year, gdpPercap, country) %>%
  filter(year==1997,
         gdpPercap>20000) %>%
  arrange(country)
```

```
## # A tibble: 27 x 3
##     year gdpPercap country
##    <int>     <dbl> <fct>
##  1  1997    26998. Australia
##  2  1997    29096. Austria
##  3  1997    20292. Bahrain
##  4  1997    27561. Belgium
##  5  1997    28955. Canada
##  6  1997    29804. Denmark
##  7  1997    23724. Finland
##  8  1997    25890. France
##  9  1997    27789. Germany
## 10  1997    28378. Hong Kong, China
## # ... with 17 more rows
```

Combine `select()` Task 1 with `arrange()` Task 3.

`filter()`

1. Only take data with population greater than 1 billion.

```
gapminder %>%
  filter(pop>1000000000)
```

```
## # A tibble: 8 x 6
##   country continent  year lifeExp        pop gdpPercap
##   <fct>   <fct>     <int>   <dbl>      <int>     <dbl>
## 1 China   Asia       1982    65.5 1000281000      962.
## 2 China   Asia       1987    67.3 1084035000     1379.
## 3 China   Asia       1992    68.7 1164970000     1656.
## 4 China   Asia       1997    70.4 1230075000     2289.
## 5 China   Asia       2002    72.0 1280400000     3119.
## 6 China   Asia       2007    73.0 1318683096     4959.
## 7 India   Asia       2002    62.9 1034172547     1747.
## 8 India   Asia       2007    64.7 1110396331     2452.
```

2. Of those, only look at data from China.

```
gapminder %>%
  filter(pop>1000000000) %>%
  filter(country=="India")
```

```
## # A tibble: 2 x 6
##   country continent  year lifeExp        pop gdpPercap
##   <fct>   <fct>     <int>   <dbl>      <int>     <dbl>
## 1 India   Asia       2002    62.9 1034172547     1747.
## 2 India   Asia       2007    64.7 1110396331     2452.
```

`mutate()`

1. Make a new variable that is `GDP` instead of `gdpPercap` (multiply `gdpPercap` by `pop`).

```r
gapminder %>%
  mutate(GDP=gdpPercap*pop)
```

```
## # A tibble: 1,704 x 7
##    country     continent  year lifeExp      pop gdpPercap          GDP
##    <fct>       <fct>     <int>   <dbl>    <int>     <dbl>        <dbl>
##  1 Afghanistan Asia       1952    28.8  8425333      779.  6567086330.
##  2 Afghanistan Asia       1957    30.3  9240934      821.  7585448670.
##  3 Afghanistan Asia       1962    32.0 10267083      853.  8758855797.
##  4 Afghanistan Asia       1967    34.0 11537966      836.  9648014150.
##  5 Afghanistan Asia       1972    36.1 13079460      740.  9678553274.
##  6 Afghanistan Asia       1977    38.4 14880372      786. 11697659231.
##  7 Afghanistan Asia       1982    39.9 12881816      978. 12598563401.
##  8 Afghanistan Asia       1987    40.8 13867957      852. 11820990309.
##  9 Afghanistan Asia       1992    41.7 16317921      649. 10595901589.
## 10 Afghanistan Asia       1997    41.8 22227415      635. 14121995875.
## # ... with 1,694 more rows
```

2. Make a new variable for `gdpPercap` that is in millions.

```r
gapminder %>%
  mutate(GDP=gdpPercap*pop) %>%
  mutate(GDPm=(GDP/1000000))
```

```
## # A tibble: 1,704 x 8
##    country     continent  year lifeExp      pop gdpPercap       GDP   GDPm
##    <fct>       <fct>     <int>   <dbl>    <int>     <dbl>     <dbl>  <dbl>
##  1 Afghanistan Asia       1952    28.8  8425333      779.   6.57e 9  6567.
##  2 Afghanistan Asia       1957    30.3  9240934      821.   7.59e 9  7585.
##  3 Afghanistan Asia       1962    32.0 10267083      853.   8.76e 9  8759.
##  4 Afghanistan Asia       1967    34.0 11537966      836.   9.65e 9  9648.
##  5 Afghanistan Asia       1972    36.1 13079460      740.   9.68e 9  9679.
##  6 Afghanistan Asia       1977    38.4 14880372      786.   1.17e10 11698.
##  7 Afghanistan Asia       1982    39.9 12881816      978.   1.26e10 12599.
##  8 Afghanistan Asia       1987    40.8 13867957      852.   1.18e10 11821.
##  9 Afghanistan Asia       1992    41.7 16317921      649.   1.06e10 10596.
## 10 Afghanistan Asia       1997    41.8 22227415      635.   1.41e10 14122.
## # ... with 1,694 more rows
```

3. Make a new population variable that is the population in millions.

```r
gapminder %>%
  mutate(popm=pop/1000000)
```

```
## # A tibble: 1,704 x 7
##    country     continent  year lifeExp      pop gdpPercap  popm
##    <fct>       <fct>     <int>   <dbl>    <int>     <dbl> <dbl>
##  1 Afghanistan Asia       1952    28.8  8425333      779.  8.43
##  2 Afghanistan Asia       1957    30.3  9240934      821.  9.24
##  3 Afghanistan Asia       1962    32.0 10267083      853. 10.3
##  4 Afghanistan Asia       1967    34.0 11537966      836. 11.5
##  5 Afghanistan Asia       1972    36.1 13079460      740. 13.1
##  6 Afghanistan Asia       1977    38.4 14880372      786. 14.9
##  7 Afghanistan Asia       1982    39.9 12881816      978. 12.9
##  8 Afghanistan Asia       1987    40.8 13867957      852. 13.9
##  9 Afghanistan Asia       1992    41.7 16317921      649. 16.3
## 10 Afghanistan Asia       1997    41.8 22227415      635. 22.2
## # ... with 1,694 more rows
```

summarize()

1. Get the average GDP per capita

```
gapminder %>%
  summarize(mean(gdpPercap))
```

```
## # A tibble: 1 x 1
##   `mean(gdpPercap)`
##               <dbl>
## 1             7215.
```

2. Get the number of observations, average, minimum, maximum, and standard deviation for GDP per capita.

```
gapminder %>%
  summarize(Obs=n(),
            Average=mean(gdpPercap),
            Minimum=min(gdpPercap),
            Maximum=max(gdpPercap),
            SD=sd(gdpPercap))
```

```
## # A tibble: 1 x 5
##     Obs Average Minimum Maximum    SD
##   <int>   <dbl>   <dbl>   <dbl> <dbl>
## 1  1704   7215.    241. 113523. 9857.
```

3. Get the average for GDP per capita, Life expectancy, and population

```
gapminder %>%
  summarize(Average_GDPcapita=mean(gdpPercap),
            Average_LE=mean(lifeExp),
            Average_pop=mean(pop))
```

```
## # A tibble: 1 x 3
##   Average_GDPcapita Average_LE Average_pop
##               <dbl>      <dbl>       <dbl>
## 1             7215.       59.5   29601212.
```

group_by()

1. Track the change in average GDP per capita over time. Hint, first group by year.

```
gapminder %>%
  group_by(year) %>%
  summarize(Average_GDPcapita=mean(gdpPercap))
```

```
## # A tibble: 12 x 2
##     year Average_GDPcapita
##    <int>             <dbl>
##  1  1952             3725.
##  2  1957             4299.
##  3  1962             4726.
##  4  1967             5484.
##  5  1972             6770.
##  6  1977             7313.
##  7  1982             7519.
##  8  1987             7901.
##  9  1992             8159.
## 10  1997             9090.
## 11  2002             9918.
## 12  2007            11680.
```

2. Get the average GDP per capita by continent.

```
gapminder %>%
  group_by(continent) %>%
  summarize(Average_GDPcapita=mean(gdpPercap))
```

```
## # A tibble: 5 x 2
##   continent Average_GDPcapita
##   <fct>                 <dbl>
## 1 Africa                2194.
## 2 Americas              7136.
## 3 Asia                  7902.
## 4 Europe               14469.
## 5 Oceania              18622.
```

3. You can group by multiple groups. Try getting the average GDP per capita by year by continent. Hint: do year first, if you do `continent` first, there are no years to group by!

```
gapminder %>%
  group_by(year, continent) %>%
  summarize(Average_GDPcapita=mean(gdpPercap))
```

```
## # A tibble: 60 x 3
## # Groups:   year [?]
##     year continent Average_GDPcapita
##    <int> <fct>                 <dbl>
##  1  1952 Africa                1253.
##  2  1952 Americas              4079.
##  3  1952 Asia                  5195.
##  4  1952 Europe                5661.
##  5  1952 Oceania              10298.
##  6  1957 Africa                1385.
##  7  1957 Americas              4616.
##  8  1957 Asia                  5788.
##  9  1957 Europe                6963.
## 10  1957 Oceania              11599.
## # ... with 50 more rows
```