The tidyverse

Install tidyverse

```
install.packages("tidyverse")
install.packages("gapminder")
library(tidyverse)
library(gapminder)
```

Vectors

```
neat_numbers <- c(1, 4, 2, 5, 7)
```

Gapminder data

```
head(gapminder) %>% kableExtra::kable()
```

country	continent	year	lifeExp	pop	gdpPercap
Afghanistan	Asia	1952	28.801	8425333	779.4453
Afghanistan	Asia	1957	30.332	9240934	820.8530
Afghanistan	Asia	1962	31.997	10267083	853.1007
Afghanistan	Asia	1967	34.020	11537966	836.1971
Afghanistan	Asia	1972	36.088	13079460	739.9811
Afghanistan	Asia	1977	38.438	14880372	786.1134

filter()

```
filter(.data = gapminder, country == "Denmark")
```

```
## # A tibble: 12 x 6
##    country continent  year lifeExp     pop gdpPercap
##    <fct>   <fct>     <int>   <dbl>   <int>     <dbl>
##  1 Denmark Europe     1952    70.8 4334000     9692.
##  2 Denmark Europe     1957    71.8 4487831    11100.
##  3 Denmark Europe     1962    72.4 4646899    13583.
##  4 Denmark Europe     1967    73.0 4838800    15937.
##  5 Denmark Europe     1972    73.5 4991596    18866.
##  6 Denmark Europe     1977    74.7 5088419    20423.
##  7 Denmark Europe     1982    74.6 5117810    21688.
##  8 Denmark Europe     1987    74.8 5127024    25116.
##  9 Denmark Europe     1992    75.3 5171393    26407.
## 10 Denmark Europe     1997    76.1 5283663    29804.
## 11 Denmark Europe     2002    77.2 5374693    32167.
## 12 Denmark Europe     2007    78.3 5468120    35278.
```

filter the data for Canada

```
filter(gapminder, country == "Canada")
```

```
## # A tibble: 12 x 6
##    country continent  year lifeExp      pop gdpPercap
##    <fct>   <fct>     <int>   <dbl>    <int>     <dbl>
##  1 Canada  Americas   1952    68.8 14785584    11367.
##  2 Canada  Americas   1957    70.0 17010154    12490.
##  3 Canada  Americas   1962    71.3 18985849    13462.
##  4 Canada  Americas   1967    72.1 20819767    16077.
##  5 Canada  Americas   1972    72.9 22284500    18971.
##  6 Canada  Americas   1977    74.2 23796400    22091.
##  7 Canada  Americas   1982    75.8 25201900    22899.
##  8 Canada  Americas   1987    76.9 26549700    26627.
##  9 Canada  Americas   1992    78.0 28523502    26343.
## 10 Canada  Americas   1997    78.6 30305843    28955.
## 11 Canada  Americas   2002    79.8 31902268    33329.
## 12 Canada  Americas   2007    80.7 33390141    36319.
```

filter all data for countries in Oceania

```
filter(gapminder, continent == "Oceania")
```

```
## # A tibble: 24 x 6
##    country   continent  year lifeExp      pop gdpPercap
##    <fct>     <fct>     <int>   <dbl>    <int>     <dbl>
##  1 Australia Oceania    1952    69.1  8691212    10040.
##  2 Australia Oceania    1957    70.3  9712569    10950.
##  3 Australia Oceania    1962    70.9 10794968    12217.
##  4 Australia Oceania    1967    71.1 11872264    14526.
##  5 Australia Oceania    1972    71.9 13177000    16789.
##  6 Australia Oceania    1977    73.5 14074100    18334.
##  7 Australia Oceania    1982    74.7 15184200    19477.
##  8 Australia Oceania    1987    76.3 16257249    21889.
##  9 Australia Oceania    1992    77.6 17481977    23425.
## 10 Australia Oceania    1997    78.8 18565243    26998.
## # ... with 14 more rows
```

filter rows where the life expectancy is greater than 82

```
filter(gapminder, lifeExp > 82)
```

```
## # A tibble: 2 x 6
##   country          continent  year lifeExp       pop gdpPercap
##   <fct>            <fct>     <int>   <dbl>     <int>     <dbl>
## 1 Hong Kong, China Asia       2007    82.2   6980412    39725.
## 2 Japan            Asia       2007    82.6 127467972    31656.
```

filter() with multiple conditions

Filter Denmark data for 2000 only

```
filter(gapminder, country == "Denmark", year > 2000)
```

```
## # A tibble: 2 x 6
##   country continent  year lifeExp     pop gdpPercap
##   <fct>   <fct>     <int>   <dbl>   <int>     <dbl>
## 1 Denmark Europe     2002    77.2 5374693    32167.
## 2 Denmark Europe     2007    78.3 5468120    35278.
```

```
filter(gapminder, country == "Denmark" & year > 2000)
```

```
## # A tibble: 2 x 6
##   country continent  year lifeExp     pop gdpPercap
##   <fct>   <fct>     <int>   <dbl>   <int>     <dbl>
## 1 Denmark Europe     2002    77.2 5374693    32167.
## 2 Denmark Europe     2007    78.3 5468120    35278.
```

Solution to filter excercise

- Canada before 1970
- Countries where life expectancy in 2007 is below 50
- Countries where life expectancy in 2007 is below 50 and are not in Africa

```
filter(gapminder, country == "Canada", year < 1970)
```

```
## # A tibble: 4 x 6
##   country continent  year lifeExp      pop gdpPercap
##   <fct>   <fct>     <int>   <dbl>    <int>     <dbl>
## 1 Canada  Americas   1952    68.8 14785584    11367.
## 2 Canada  Americas   1957    70.0 17010154    12490.
## 3 Canada  Americas   1962    71.3 18985849    13462.
## 4 Canada  Americas   1967    72.1 20819767    16077.
```

```
filter(gapminder, year == 2007, lifeExp < 50)
```

```
## # A tibble: 19 x 6
##    country                  continent  year lifeExp       pop gdpPercap
##    <fct>                    <fct>     <int>   <dbl>     <int>     <dbl>
##  1 Afghanistan              Asia       2007    43.8  31889923      975.
##  2 Angola                   Africa     2007    42.7  12420476     4797.
##  3 Burundi                  Africa     2007    49.6   8390505      430.
##  4 Central African Republic Africa     2007    44.7   4369038      706.
##  5 Congo, Dem. Rep.         Africa     2007    46.5  64606759      278.
##  6 Cote d'Ivoire            Africa     2007    48.3  18013409     1545.
##  7 Guinea-Bissau            Africa     2007    46.4   1472041      579.
##  8 Lesotho                  Africa     2007    42.6   2012649     1569.
##  9 Liberia                  Africa     2007    45.7   3193942      415.
## 10 Malawi                   Africa     2007    48.3  13327079      759.
## 11 Mozambique               Africa     2007    42.1  19951656      824.
## 12 Nigeria                  Africa     2007    46.9 135031164     2014.
## 13 Rwanda                   Africa     2007    46.2   8860588      863.
## 14 Sierra Leone             Africa     2007    42.6   6144562      863.
## 15 Somalia                  Africa     2007    48.2   9118773      926.
## 16 South Africa             Africa     2007    49.3  43997828     9270.
## 17 Swaziland                Africa     2007    39.6   1133066     4513.
## 18 Zambia                   Africa     2007    42.4  11746035     1271.
## 19 Zimbabwe                 Africa     2007    43.5  12311143      470.
```

```
filter(gapminder, year == 2007, lifeExp < 50, continent != "Africa")
```

```
## # A tibble: 1 x 6
##   country     continent  year lifeExp      pop gdpPercap
##   <fct>       <fct>     <int>   <dbl>    <int>     <dbl>
## 1 Afghanistan Asia       2007    43.8 31889923      975.
```

mutate()

```
mutate(gapminder, 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
```

```
mutate(gapminder, gdp = gdpPercap * pop, pop_mil = round(pop/1e+06))
```

```
## # A tibble: 1,704 x 8
##    country     continent  year lifeExp      pop gdpPercap          gdp pop_mil
##    <fct>       <fct>     <int>   <dbl>    <int>     <dbl>        <dbl>   <dbl>
##  1 Afghanistan Asia       1952    28.8  8425333      779.  6567086330.       8
##  2 Afghanistan Asia       1957    30.3  9240934      821.  7585448670.       9
##  3 Afghanistan Asia       1962    32.0 10267083      853.  8758855797.      10
##  4 Afghanistan Asia       1967    34.0 11537966      836.  9648014150.      12
##  5 Afghanistan Asia       1972    36.1 13079460      740.  9678553274.      13
##  6 Afghanistan Asia       1977    38.4 14880372      786. 11697659231.      15
##  7 Afghanistan Asia       1982    39.9 12881816      978. 12598563401.      13
##  8 Afghanistan Asia       1987    40.8 13867957      852. 11820990309.      14
##  9 Afghanistan Asia       1992    41.7 16317921      649. 10595901589.      16
## 10 Afghanistan Asia       1997    41.8 22227415      635. 14121995875.      22
## # ... with 1,694 more rows
```

mutate() and ifelse()

```
mutate(gapminder, after_1960 = ifelse(year > 1960, TRUE, FALSE))
```

```
## # A tibble: 1,704 x 7
##    country     continent  year lifeExp      pop gdpPercap after_1960
##    <fct>       <fct>     <int>   <dbl>    <int>     <dbl> <lgl>
##  1 Afghanistan Asia       1952    28.8  8425333      779. FALSE
##  2 Afghanistan Asia       1957    30.3  9240934      821. FALSE
##  3 Afghanistan Asia       1962    32.0 10267083      853. TRUE
##  4 Afghanistan Asia       1967    34.0 11537966      836. TRUE
##  5 Afghanistan Asia       1972    36.1 13079460      740. TRUE
##  6 Afghanistan Asia       1977    38.4 14880372      786. TRUE
```

```
##  7 Afghanistan Asia       1982    39.9 12881816      978. TRUE
##  8 Afghanistan Asia       1987    40.8 13867957      852. TRUE
##  9 Afghanistan Asia       1992    41.7 16317921      649. TRUE
## 10 Afghanistan Asia       1997    41.8 22227415      635. TRUE
## # ... with 1,694 more rows
```

```
mutate(gapminder, after_1960 = ifelse(year > 1960, "After 1960", "Before 1960"))
```

```
## # A tibble: 1,704 x 7
##    country     continent  year lifeExp      pop gdpPercap after_1960
##    <fct>       <fct>     <int>   <dbl>    <int>     <dbl> <chr>
##  1 Afghanistan Asia       1952    28.8  8425333      779. Before 1960
##  2 Afghanistan Asia       1957    30.3  9240934      821. Before 1960
##  3 Afghanistan Asia       1962    32.0 10267083      853. After 1960
##  4 Afghanistan Asia       1967    34.0 11537966      836. After 1960
##  5 Afghanistan Asia       1972    36.1 13079460      740. After 1960
##  6 Afghanistan Asia       1977    38.4 14880372      786. After 1960
##  7 Afghanistan Asia       1982    39.9 12881816      978. After 1960
##  8 Afghanistan Asia       1987    40.8 13867957      852. After 1960
##  9 Afghanistan Asia       1992    41.7 16317921      649. After 1960
## 10 Afghanistan Asia       1997    41.8 22227415      635. After 1960
## # ... with 1,694 more rows
```

- Add an `africa` column that is TRUE if the country is on the African continent
- Add a column for logged GDP per capita (hint: use log())
- Add an `africa_asia` column that says "Africa or Asia" if the country is in Africa or Asia, and "Not Africa or Asia" if it's not

```
mutate(gapminder, africa = ifelse(continent == "Africa", TRUE, FALSE))
```

```
mutate(gapminder, log_gdpPercap = log(gdpPercap))
```

```
mutate(gapminder, africa_asia = ifelse(continent %in% c("Africa", "Asia"), "Africa or Asia",
    "Not Africa or Asia"))
```

creating new dataframes

```
gapminder_2002 <- filter(gapminder, year == 2002)

gapminder_2002_log <- mutate(gapminder_2002, log_gdpPercap = log(gdpPercap))
```

```
gapminder_2002b <- gapminder %>% filter(year == 2002) %>% mutate(log_gdpPercap = log(gdpPercap))

gapminder_2002b
```

```
## # A tibble: 142 x 7
##    country     continent  year lifeExp       pop gdpPercap log_gdpPercap
##    <fct>       <fct>     <int>   <dbl>     <int>     <dbl>         <dbl>
##  1 Afghanistan Asia       2002    42.1  25268405      727.          6.59
##  2 Albania     Europe     2002    75.7   3508512     4604.          8.43
##  3 Algeria     Africa     2002    71.0  31287142     5288.          8.57
##  4 Angola      Africa     2002    41.0  10866106     2773.          7.93
##  5 Argentina   Americas   2002    74.3  38331121     8798.          9.08
```

```
##  6 Australia   Oceania    2002    80.4  19546792    30688.         10.3
##  7 Austria     Europe     2002    79.0   8148312    32418.         10.4
##  8 Bahrain     Asia       2002    74.8    656397    23404.         10.1
##  9 Bangladesh  Asia       2002    62.0 135656790     1136.          7.04
## 10 Belgium     Europe     2002    78.3  10311970    30486.         10.3
## # ... with 132 more rows
```

summarize()

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

```
## # A tibble: 1 x 1
##   mean_life
##       <dbl>
## 1      59.5
```

```
gapminder %>% summarize(mean_life = mean(lifeExp), min_life = min(lifeExp))
```

```
## # A tibble: 1 x 2
##   mean_life min_life
##       <dbl>    <dbl>
## 1      59.5     23.6
```

```
gapminder %>% summarize(first = min(year), last = max(year), num_rows = n(), num_unique = n_distinct(co
```

groupby() and summarize()

```
gapminder %>% group_by(continent) %>% summarize(n_countries = n_distinct(country))
```

```
## # A tibble: 5 x 2
##   continent n_countries
## * <fct>           <int>
## 1 Africa             52
## 2 Americas           25
## 3 Asia               33
## 4 Europe             30
## 5 Oceania             2
```

```
## solution

gapminder %>% group_by(continent) %>% summarize(min_le = min(lifeExp), max_le = max(lifeExp),
    med_le = median(lifeExp))



gapminder %>% filter(year == 2007) %>% group_by(continent) %>% summarize(min_le = min(lifeExp),
    max_le = max(lifeExp), med_le = median(lifeExp))
```