Lab 3- Data Wrangling - Questions

Bella Fascendini

2024-09-25

Lab 3 - Gapminder Data Wrangling Lab Assignment

Using the Gapminder dataset, complete the following tasks. Use tidyverse and dplyr functions and pipes where appropriate. Remember to load the necessary libraries and the Gapminder dataset before starting.

1) Find all countries in Asia with a life expectancy greater than 75 years in 2007.

```
head(gapminder)
```

```
## # A tibble: 6 x 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.
```

```
asia_75plus_2007 <- gapminder %>%
  filter(year == 2007, continent == "Asia", lifeExp > 75) %>%
  select(country, lifeExp)

asia_75plus_2007
```

```
## # A tibble: 9 x 2
##   country          lifeExp
##   <fct>              <dbl>
## 1 Bahrain             75.6
## 2 Hong Kong, China    82.2
## 3 Israel              80.7
## 4 Japan               82.6
## 5 Korea, Rep.         78.6
## 6 Kuwait              77.6
## 7 Oman                75.6
## 8 Singapore           80.0
## 9 Taiwan              78.4
```

2) List the top 5 countries with the highest GDP per capita in 2007, in descending order.

```
top_5_gdp <- gapminder %>%
  filter(year == 2007) %>%
  arrange(desc(gdpPercap)) %>%
  select(country, gdpPercap) %>%
  slice(1:5)
top_5_gdp
```

```
## # A tibble: 5 x 2
##   country       gdpPercap
##   <fct>             <dbl>
## 1 Norway            49357.
## 2 Kuwait            47307.
## 3 Singapore         47143.
## 4 United States     42952.
## 5 Ireland           40676.
```

3) Create a new dataframe with only the country, continent, and GDP per capita for the year 2007.

```
df_2007 <- gapminder %>%
  filter(year == 2007) %>%
  select(country, continent, gdpPercap)

df_2007
```

```
## # A tibble: 142 x 3
##    country     continent gdpPercap
##    <fct>       <fct>         <dbl>
##  1 Afghanistan Asia           975.
##  2 Albania     Europe        5937.
##  3 Algeria     Africa        6223.
##  4 Angola      Africa        4797.
##  5 Argentina   Americas     12779.
##  6 Australia   Oceania      34435.
##  7 Austria     Europe       36126.
##  8 Bahrain     Asia         29796.
##  9 Bangladesh  Asia          1391.
## 10 Belgium     Europe       33693.
## # i 132 more rows
```

4) Add a new column called "gdp" that calculates the total GDP (GDP per capita * population) for each country in 2007.

```
gdp_total <- df_2007 %>%
  mutate(gdp = gdpPercap * gapminder$pop[gapminder$year == 2007])

gdp_total
```

```
## # A tibble: 142 x 4
##    country     continent gdpPercap          gdp
##    <fct>       <fct>         <dbl>        <dbl>
```

```
##  1 Afghanistan Asia              975.  31079291949.
##  2 Albania      Europe          5937.  21376411360.
##  3 Algeria      Africa          6223. 207444851958.
##  4 Angola       Africa          4797.  59583895818.
##  5 Argentina    Americas       12779. 515033625357.
##  6 Australia    Oceania        34435. 703658358894.
##  7 Austria      Europe         36126. 296229400691.
##  8 Bahrain      Asia           29796.  21112675360.
##  9 Bangladesh   Asia            1391. 209311822134.
## 10 Belgium      Europe         33693. 350141166520.
## # i 132 more rows
```

5) Create two separate dataframes: one for countries in Europe and one for countries in Africa, both for
   the year 2007. Then bind these dataframes together.

```r
europe_2007 <- gapminder %>%
  filter(year == 2007, continent == "Europe")

africa_2007 <- gapminder %>%
  filter(year == 2007, continent == "Africa")

europe_africa_combined <- bind_rows(europe_2007, africa_2007)

europe_africa_combined
```

```
## # A tibble: 82 x 6
##    country                continent  year lifeExp       pop gdpPercap
##    <fct>                  <fct>     <int>   <dbl>     <int>     <dbl>
##  1 Albania                Europe     2007    76.4   3600523     5937.
##  2 Austria                Europe     2007    79.8   8199783    36126.
##  3 Belgium                Europe     2007    79.4  10392226    33693.
##  4 Bosnia and Herzegovina Europe     2007    74.9   4552198     7446.
##  5 Bulgaria               Europe     2007    73.0   7322858    10681.
##  6 Croatia                Europe     2007    75.7   4493312    14619.
##  7 Czech Republic         Europe     2007    76.5  10228744    22833.
##  8 Denmark                Europe     2007    78.3   5468120    35278.
##  9 Finland                Europe     2007    79.3   5238460    33207.
## 10 France                 Europe     2007    80.7  61083916    30470.
## # i 72 more rows
```

6) Calculate the average life expectancy and total population for each continent in 2007.

```r
summary_2007 <- gapminder %>%
  filter(year == 2007) %>%
  group_by(continent) %>%
  summarize(
    average_life_expectancy = mean(lifeExp),
    total_population = sum(pop)
  )

summary_2007
```

```
## # A tibble: 5 x 3
##   continent average_life_expectancy total_population
##   <fct>                       <dbl>            <dbl>
## 1 Africa                       54.8        929539692
## 2 Americas                     73.6        898871184
## 3 Asia                         70.7       3811953827
## 4 Europe                       77.6        586098529
## 5 Oceania                      80.7         24549947
```

7) Create a wide format dataframe that shows the population for each country across all life expectancies as columns.

```
population_wide <- gapminder %>%
  #filter(year == 2007) %>%
  select(country, pop, year) %>%
  pivot_wider(names_from = country,
              values_from = pop,
              names_prefix = "pop_")
population_wide
```

```
## # A tibble: 12 x 143
##     year pop_Afghanistan pop_Albania pop_Algeria pop_Angola pop_Argentina
##    <int>           <int>       <int>       <int>      <int>         <int>
## 1   1952         8425333     1282697     9279525    4232095      17876956
## 2   1957         9240934     1476505    10270856    4561361      19610538
## 3   1962        10267083     1728137    11000948    4826015      21283783
## 4   1967        11537966     1984060    12760499    5247469      22934225
## 5   1972        13079460     2263554    14760787    5894858      24779799
## 6   1977        14880372     2509048    17152804    6162675      26983828
## 7   1982        12881816     2780097    20033753    7016384      29341374
## 8   1987        13867957     3075321    23254956    7874230      31620918
## 9   1992        16317921     3326498    26298373    8735988      33958947
## 10  1997        22227415     3428038    29072015    9875024      36203463
## 11  2002        25268405     3508512    31287142   10866106      38331121
## 12  2007        31889923     3600523    33333216   12420476      40301927
## # i 137 more variables: pop_Australia <int>, pop_Austria <int>,
## #   pop_Bahrain <int>, pop_Bangladesh <int>, pop_Belgium <int>,
## #   pop_Benin <int>, pop_Bolivia <int>, `pop_Bosnia and Herzegovina` <int>,
## #   pop_Botswana <int>, pop_Brazil <int>, pop_Bulgaria <int>,
## #   `pop_Burkina Faso` <int>, pop_Burundi <int>, pop_Cambodia <int>,
## #   pop_Cameroon <int>, pop_Canada <int>, `pop_Central African Republic` <int>,
## #   pop_Chad <int>, pop_Chile <int>, pop_China <int>, pop_Colombia <int>, ...
```