Wk 6 Data Transformation: World Population

Dirk Hartog

2023-10-08

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

We will looking at a data set containing global population counts for 234 countries or territories.

Read in the untidy .csv file from github

```
url <- "https://raw.githubusercontent.com/D-hartog/DATA607/main/PROJECT2/worldpop_untidy.csv"

world_pop <- read_csv(url)
```

```
## Rows: 234 Columns: 17
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (4): CCA3, Country/Territory, Capital, Continent
## dbl (13): Rank, 2022 Population, 2020 Population, 2015 Population, 2010 Popu...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(world_pop)
```

```
## # A tibble: 6 x 17
##    Rank CCA3  `Country/Territory` Capital          Continent `2022 Population`
##   <dbl> <chr> <chr>               <chr>            <chr>                  <dbl>
## 1    36 AFG   Afghanistan         Kabul            Asia                41128771
## 2   138 ALB   Albania             Tirana           Europe               2842321
```

```
## 3    34 DZA   Algeria             Algiers          Africa             44903225
## 4   213 ASM   American Samoa      Pago Pago        Oceania               44273
## 5   203 AND   Andorra             Andorra la Vella Europe                79824
## 6    42 AGO   Angola              Luanda           Africa             35588987
## # i 11 more variables: `2020 Population` <dbl>, `2015 Population` <dbl>,
## #   `2010 Population` <dbl>, `2000 Population` <dbl>, `1990 Population` <dbl>,
## #   `1980 Population` <dbl>, `1970 Population` <dbl>, `Area (km²)` <dbl>,
## #   `Density (per km²)` <dbl>, `Growth Rate` <dbl>,
## #   `World Population Percentage` <dbl>
```

```
glimpse(world_pop)
```

```
## Rows: 234
## Columns: 17
## $ Rank                          <dbl> 36, 138, 34, 213, 203, 42, 224, 201, 33,~
## $ CCA3                          <chr> "AFG", "ALB", "DZA", "ASM", "AND", "AGO"~
## $ `Country/Territory`           <chr> "Afghanistan", "Albania", "Algeria", "Am~
## $ Capital                       <chr> "Kabul", "Tirana", "Algiers", "Pago Pago~
## $ Continent                     <chr> "Asia", "Europe", "Africa", "Oceania", "~
## $ `2022 Population`             <dbl> 41128771, 2842321, 44903225, 44273, 7982~
## $ `2020 Population`             <dbl> 38972230, 2866849, 43451666, 46189, 7770~
## $ `2015 Population`             <dbl> 33753499, 2882481, 39543154, 51368, 7174~
## $ `2010 Population`             <dbl> 28189672, 2913399, 35856344, 54849, 7151~
## $ `2000 Population`             <dbl> 19542982, 3182021, 30774621, 58230, 6609~
## $ `1990 Population`             <dbl> 10694796, 3295066, 25518074, 47818, 5356~
## $ `1980 Population`             <dbl> 12486631, 2941651, 18739378, 32886, 3561~
## $ `1970 Population`             <dbl> 10752971, 2324731, 13795915, 27075, 1986~
## $ `Area (km²)`                  <dbl> 652230, 28748, 2381741, 199, 468, 124670~
## $ `Density (per km²)`           <dbl> 63.0587, 98.8702, 18.8531, 222.4774, 170~
## $ `Growth Rate`                 <dbl> 1.0257, 0.9957, 1.0164, 0.9831, 1.0100, ~
## $ `World Population Percentage` <dbl> 0.52, 0.04, 0.56, 0.00, 0.00, 0.45, 0.00~
```

CLEANING THE DATA

1. First I want to change the column names for later transformtion of the year columns .

```
colnames(world_pop)[c(1:17)] <- c("RANK","CCAS", "COUNTRY_TERR", "CAPITAL","CONTINENT", "2022", "2020",
```

TIDY/TRANSFORMING THE DATA: For the analysis I want to compare growth rates by year as well as among continents. In order to make do this analysis we will need to created an observation for each country and year.

1. I don't think that there is much transformation that needs to be done to the data execpt taking the year columns pivoting those columns to rows.

```
world_long <- world_pop %>%
  pivot_longer(
    cols = "2022":"1970",
    names_to = "YEAR",
    values_to = "POPULATION"
```

```
  )

glimpse(world_long)
```

```
## Rows: 1,872
## Columns: 11
## $ RANK          <dbl> 36, 36, 36, 36, 36, 36, 36, 36, 138, 138, 138, 138, 138,~
## $ CCAS          <chr> "AFG", "AFG", "AFG", "AFG", "AFG", "AFG", "AFG", "AFG", ~
## $ COUNTRY_TERR  <chr> "Afghanistan", "Afghanistan", "Afghanistan", "Afghanista~
## $ CAPITAL       <chr> "Kabul", "Kabul", "Kabul", "Kabul", "Kabul", "Kabul", "K~
## $ CONTINENT     <chr> "Asia", "Asia", "Asia", "Asia", "Asia", "Asia", "Asia", ~
## $ AREA          <dbl> 652230, 652230, 652230, 652230, 652230, 652230, 652230, ~
## $ DENSITY       <dbl> 63.0587, 63.0587, 63.0587, 63.0587, 63.0587, 63.0587, 63~
## $ GROWTH_RATE   <dbl> 1.0257, 1.0257, 1.0257, 1.0257, 1.0257, 1.0257, 1.0257, ~
## $ WORLD_POP_PCT <dbl> 0.52, 0.52, 0.52, 0.52, 0.52, 0.52, 0.52, 0.52, 0.04, 0.~
## $ YEAR          <chr> "2022", "2020", "2015", "2010", "2000", "1990", "1980", ~
## $ POPULATION    <dbl> 41128771, 38972230, 33753499, 28189672, 19542982, 106947~
```

```
write.csv(world_long,file='/Users/dirkhartog/Desktop/CUNY_MSDS/DATA_607/PROJECT2/world_pop/worldpop_tid
```

DATA ANALYSIS AND VISUALIZATIONS

1. Finding the total, min and max populations for each year and each and continent

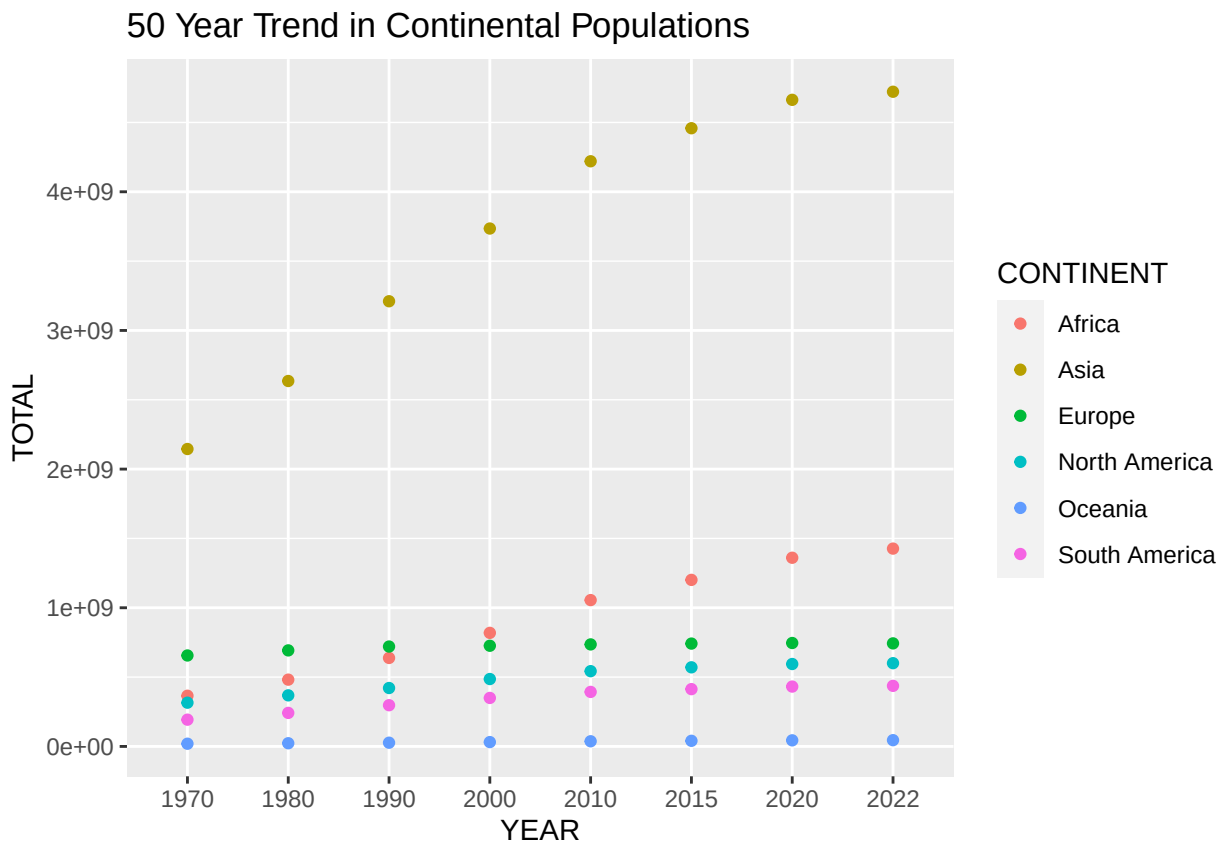
```
world_long %>%
  group_by(CONTINENT, YEAR) %>%
  summarize(TOTAL = sum(POPULATION, na.rm = TRUE),
         MAX = max(POPULATION, na.rm = TRUE),
         MIN = min(POPULATION, na.rm = TRUE))
```

```
## `summarise()` has grouped output by 'CONTINENT'. You can override using the
## `.groups` argument.
```

```
## # A tibble: 48 x 5
## # Groups:   CONTINENT [6]
##    CONTINENT YEAR       TOTAL       MAX    MIN
##    <chr>     <chr>      <dbl>     <dbl>  <dbl>
##  1 Africa    1970   365444348  55569264  35383
##  2 Africa    1980   481536377  72951439  52233
##  3 Africa    1990   638150629  95214257  71057
##  4 Africa    2000   818946032 122851984  80060
##  5 Africa    2010  1055228072 160952853  92409
##  6 Africa    2015  1201102442 183995785  99240
##  7 Africa    2020  1360671810 208327405 105530
##  8 Africa    2022  1426730932 218541212 107118
##  9 Asia      1970  2144906290 822534450 118007
## 10 Asia      1980  2635334228 982372466 164887
## # i 38 more rows
```

```
world_long %>% group_by(CONTINENT, YEAR) %>%
  summarise(TOTAL = sum(POPULATION, na.rm = TRUE)) %>%
  ggplot(aes(x = YEAR, y = TOTAL)) +
  geom_point(aes(color = CONTINENT)) +
  ggtitle("50 Year Trend in Continental Populations")
```

```
## `summarise()` has grouped output by 'CONTINENT'. You can override using the
## `.groups` argument.
```

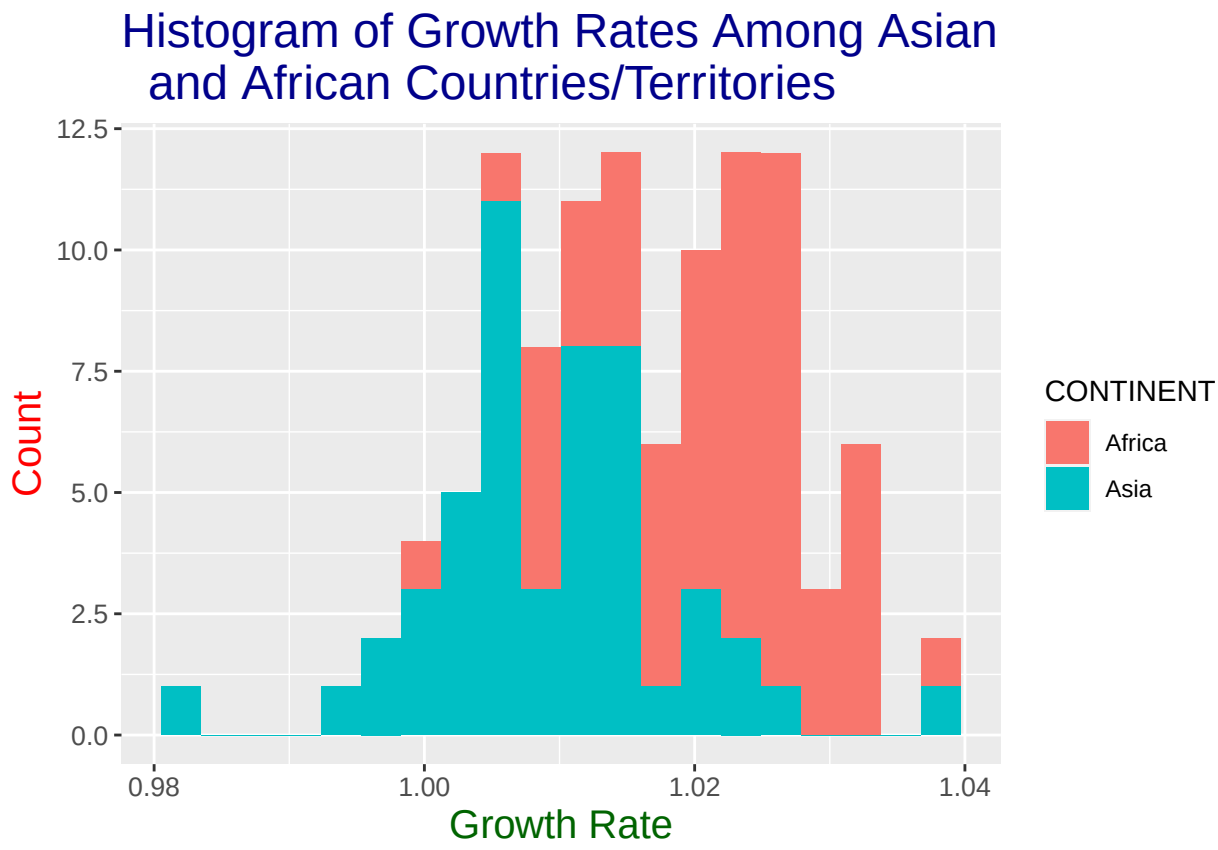


2. We can now pull out information for a given year and compare attributes, in this case among continents. Looking at some statistics and trends in growth rates from Asian and African countries in 2022

```
world_long %>%  filter(YEAR == "2022") %>%
  group_by(CONTINENT) %>%
  summarise(Average_gr = mean(GROWTH_RATE, na.rm = TRUE))
```

```
## # A tibble: 6 x 2
##   CONTINENT     Average_gr
##   <chr>              <dbl>
## 1 Africa              1.02
## 2 Asia                1.01
## 3 Europe              1.00
## 4 North America       1.00
## 5 Oceania             1.01
## 6 South America       1.01
```

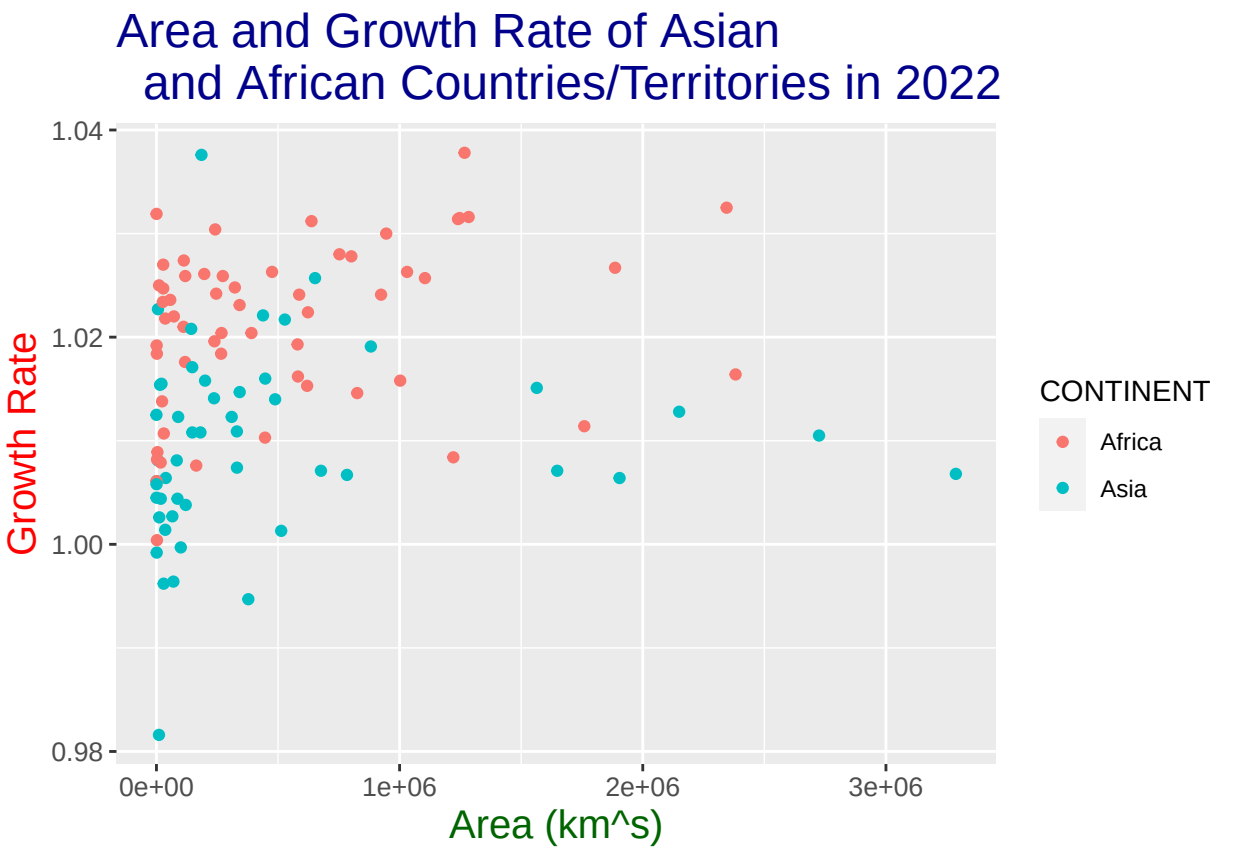
```r
world_long %>%
  filter(YEAR == "2022" & CONTINENT %in% c("Asia", "Africa")) %>%
  ggplot(aes(x = GROWTH_RATE)) +
  geom_histogram(bins = 20, aes(fill = CONTINENT)) +
  ggtitle("Histogram of Growth Rates Among Asian
  and African Countries/Territories") +
  ylab("Count") +
  xlab("Growth Rate") +
  theme(axis.title.x = element_text(color="darkgreen",size=15),
        axis.title.y = element_text(color="red", size=15),
        axis.text.x = element_text(size=10),
        axis.text.y = element_text(size=10),
        plot.title = element_text(color="darkblue",
                                  size=18))
```



3. Looking at any relationships in area size less than 5 million (km^2) and growth rate

```r
world_long %>% filter(YEAR == "2022" & AREA < 5000000 & CONTINENT %in% c("Asia", "Africa")) %>%
  ggplot(aes(x = AREA, y = GROWTH_RATE, color = CONTINENT)) +
  geom_point() +
  ggtitle("Area and Growth Rate of Asian
  and African Countries/Territories in 2022") +
  ylab("Growth Rate") +
  xlab("Area (km^s)") +
  theme(axis.title.x = element_text(color="darkgreen",size=15),
```

```
        axis.title.y = element_text(color="red", size=15),
        axis.text.x = element_text(size=10),
        axis.text.y = element_text(size=10),
        plot.title = element_text(color="darkblue",
                                  size=18))
```



CONCLUSIONS It is clear that Asian and African countries have seen a larger trend in the growth of their populations over the past 50 years. Despite this it does seem that currently average growth rates across the globe are pretty similar between 1.002 - 1.02. This growth rate might seem small but when talking about populations, a 1% grwoth rate is still a lot of people!