DATA 607, Project Two: Data Transformation - World Population

Kevin Kirby

2024-09-29

Overview

This is based on the world population trends dataset submitted by NOM. The table has population snapshots every decade or so from 1970 to 2022.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v stringr   1.5.1
## v lubridate 1.9.3     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.1
## v readr     2.1.5

## -- 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 error
```

```
library(readr)

world_population_nom <- read.csv("https://storage.googleapis.com/data_science_masters_files/2024_fall/da
```

Tidying the data

The main thing was dropping off columns I didn't need, like country code or a separate ranking column.I rewired the table so that the row number is the ranking in population. I changed field names to get away from dot syntax, which shouldn't be used for fields within a table.

```
tidy_world_data <- world_population_nom
tidy_world_data <- tidy_world_data[order(tidy_world_data$Rank), ]

tidy_world_data$Rank <- NULL
tidy_world_data$CCA3 <- NULL
```

```r
tidy_world_data <- tidy_world_data %>%
  rename(
    country_territory = Country.Territory,
    capital = Capital,
    continent = Continent,
    pop_2022 = X2022.Population,
    pop_2020 = X2020.Population,
    pop_2015 = X2015.Population,
    pop_2010 = X2010.Population,
    pop_2000 = X2000.Population,
    pop_1990 = X1990.Population,
    pop_1980 = X1980.Population,
    pop_1970 = X1970.Population,
    area_km = Area..km..,
    density_per_km = Density..per.km..,
    growth_rate = Growth.Rate,
    world_pop_percentage = World.Population.Percentage
  )

rownames(tidy_world_data) <- NULL
```

Analysis

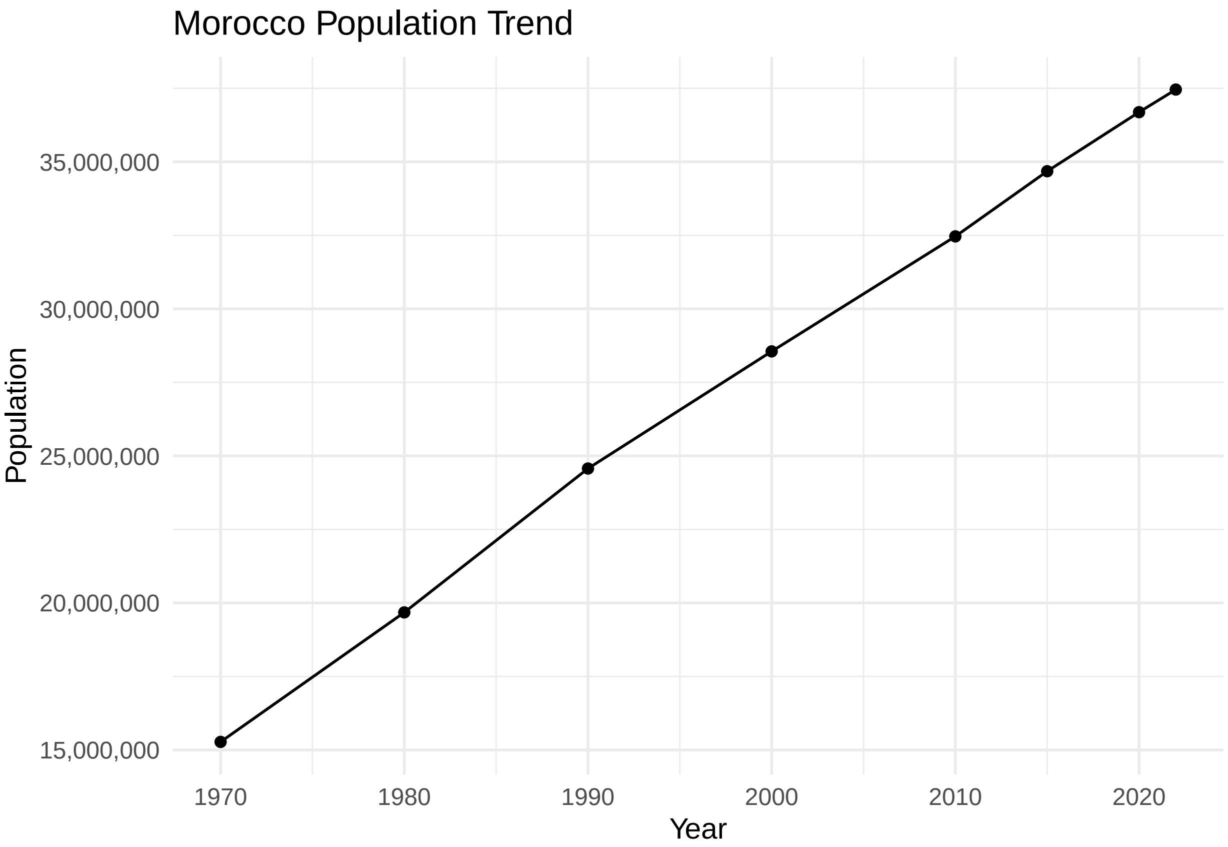
The discussion board prompt from NOM stated: "With this dataset, analyzing population dynamics, both historically and geographically can be conducted. Trends over time, comparing regions and countries, and exploring relationships between different variables such as growth rates and population density."

The first chart shows overall population growth in Morocco between 1970 and 2022. The second chart shows a log 10 relationship between population density and and then growth rate. In general I was looking for initial evidence showing that a country experiencing increased density is more likely to experience population growth.

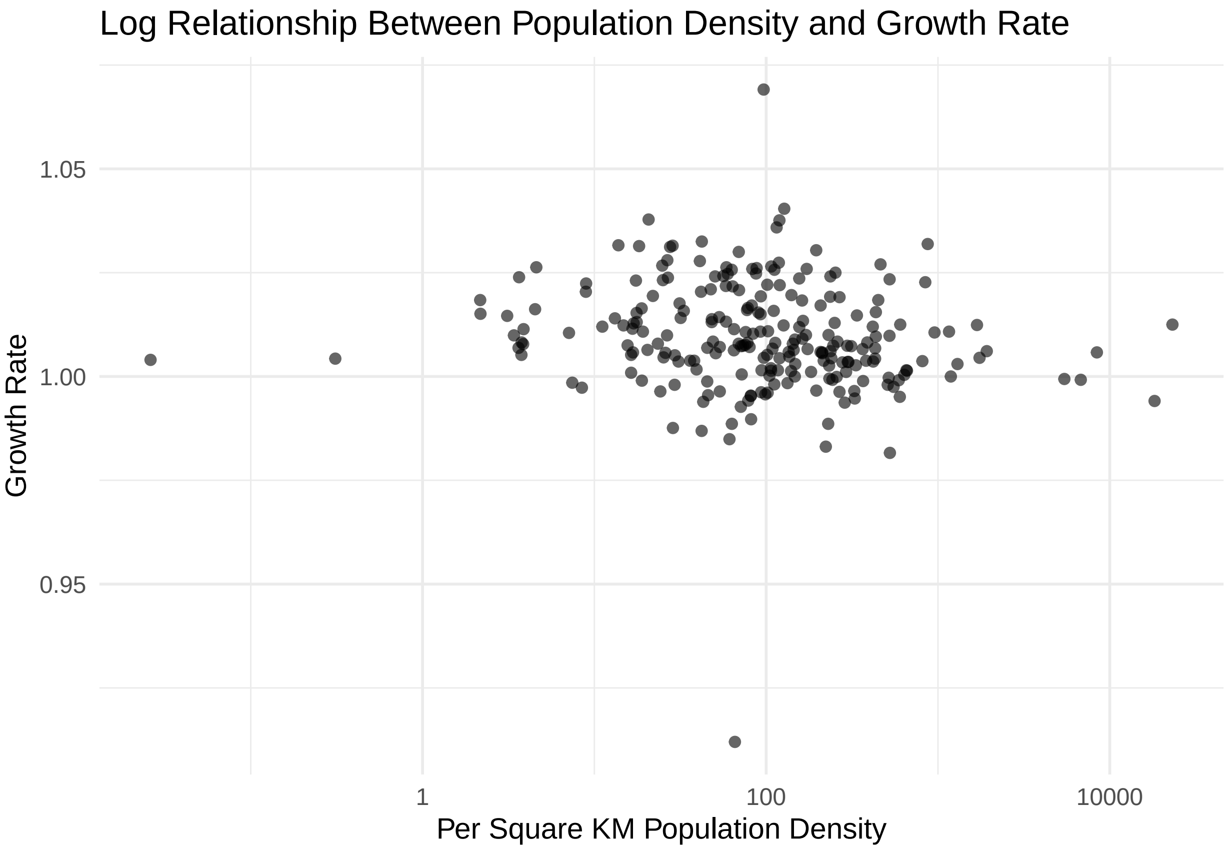
```r
country_data <- subset(tidy_world_data, country_territory == "Morocco")
years <- c(1970, 1980, 1990, 2000, 2010, 2015, 2020, 2022)
population_values <- as.numeric(country_data[1, c("pop_1970", "pop_1980", "pop_1990", "pop_2000", "pop_2

population_trend <- data.frame(year = years, population = population_values)

ggplot(population_trend, aes(x = year, y = population)) +
  geom_line() +
  geom_point() +
  ggtitle("Morocco Population Trend") +
  xlab("Year") +
  ylab("Population") +
  scale_y_continuous(labels = scales::comma) +
  theme_minimal()
```



```
ggplot(tidy_world_data, aes(x = density_per_km, y = growth_rate)) +
  geom_point(alpha = 0.6) +
  ggtitle("Log Relationship Between Population Density and Growth Rate") +
  xlab("Per Square KM Population Density") +
  ylab("Growth Rate") +
  theme_minimal() +
  scale_x_log10()
```



Conclusion

There isn't much of a correlation between population density and growth rate. Those plotted points are scattered rather chaotically and don't fit a particular narrative. More people do beget more people so some sort of density is required, although it may be far lower than I thought.