Project2

Nicholas Kunze

2024-03-24

Summary

In this project, I showcase preparing datasets for downstream analysis work; this specifically deals with wide datasets and getting them ready by making them follow the rules of a tidy dataset. For this project, I chose my own discussion item, world population data and projects.

World Population History and Projections

Get table from MySQL server and get head of new DataFrame.

```
azuredb = dbConnect(MySQL(), user=params$dbuser, password=params$dbpass, dbname=params$dbname, host=para

pop <- dbGetQuery(azuredb,
  "SELECT
      *
   FROM
      world_population_data;")
head(pop)
```

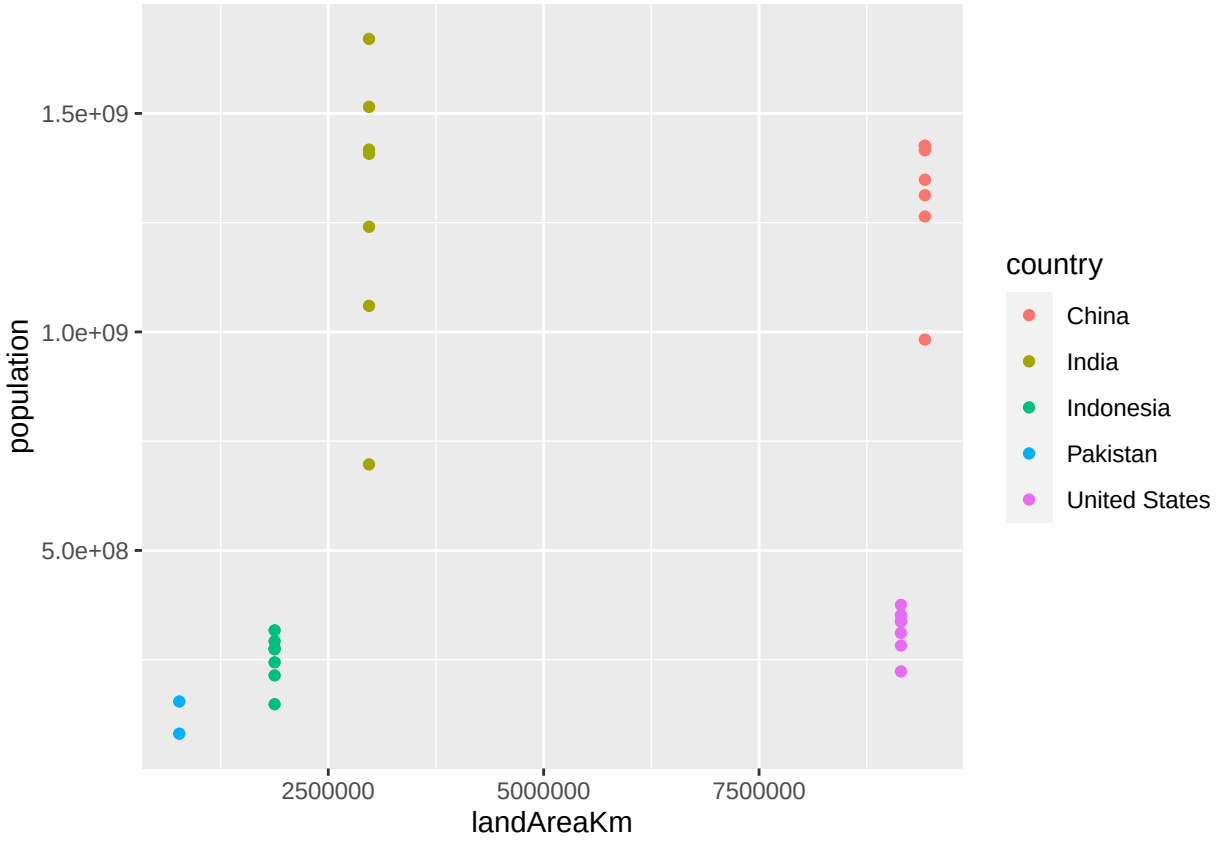
```
##   MyUnknownColumn Rank       country country_code      1980       2000
## 1               0    1         China          CHN 982372466 1264099069
## 2               1    2         India          IND 696828385 1059633675
## 3               2    3 United States          USA 223140018  282398554
## 4               3    4     Indonesia          IDN 148177096  214072421
## 5               4    5      Pakistan          PAK  80624057  154369924
## 6               5    6       Nigeria          NGA  72951439  122851984
##         2010       2021       2022       2030       2050    area landAreaKm
## 1 1348191368 1425893465 1425887337 1415605906 1312636325 9706961    9424703
## 2 1240613620 1407563842 1417173173 1514994080 1670490596 3287590    2973190
## 3  311182845  336997624  338289857  352162301  375391963 9372610    9147420
## 4  244016173  273753191  275501339  292150100  317225213 1904569    1877519
## 5  194454498  231402117  235824862  274029836  367808468  881912     770880
## 6  160952853  213401323  218541212  262580425  377459883  923768     910770
##   growthRate worldPercentage  density
## 1     0.0000          0.1788 151.2926
## 2     0.0068          0.1777 476.6507
## 3     0.0038          0.0424  36.9820
## 4     0.0064          0.0345 146.7369
## 5     0.0191          0.0296 305.9164
## 6     0.0241          0.0274 239.9521
```

As we can see, this is a wide df showing information for a single country over multiple years with a few unneeded columns. We'd like each row to be a single observation of population for a country. We should gather up these years that are columns, as they are values for 'year', not variable names. Then we can just select the variables we need for analysis.

```
pop <- pop %>%
  pivot_longer(cols = `1980`:`2050`, names_to = "year", values_to = "population")
pop <- pop[,c("country","year","population","landAreaKm","growthRate")]
head(pop)
```

```
## # A tibble: 6 x 5
##   country year  population landAreaKm growthRate
##   <chr>   <chr>      <dbl>      <dbl>      <dbl>
## 1 China   1980   982372466   9424703.          0
## 2 China   2000  1264099069   9424703.          0
## 3 China   2010  1348191368   9424703.          0
## 4 China   2021  1425893465   9424703.          0
## 5 China   2022  1425887337   9424703.          0
## 6 China   2030  1415605906   9424703.          0
```

```
pop_top <- head(pop,30)
ggplot(data = pop_top, mapping = aes(x = landAreaKm, y = population, color = country)) +
  geom_point()
```



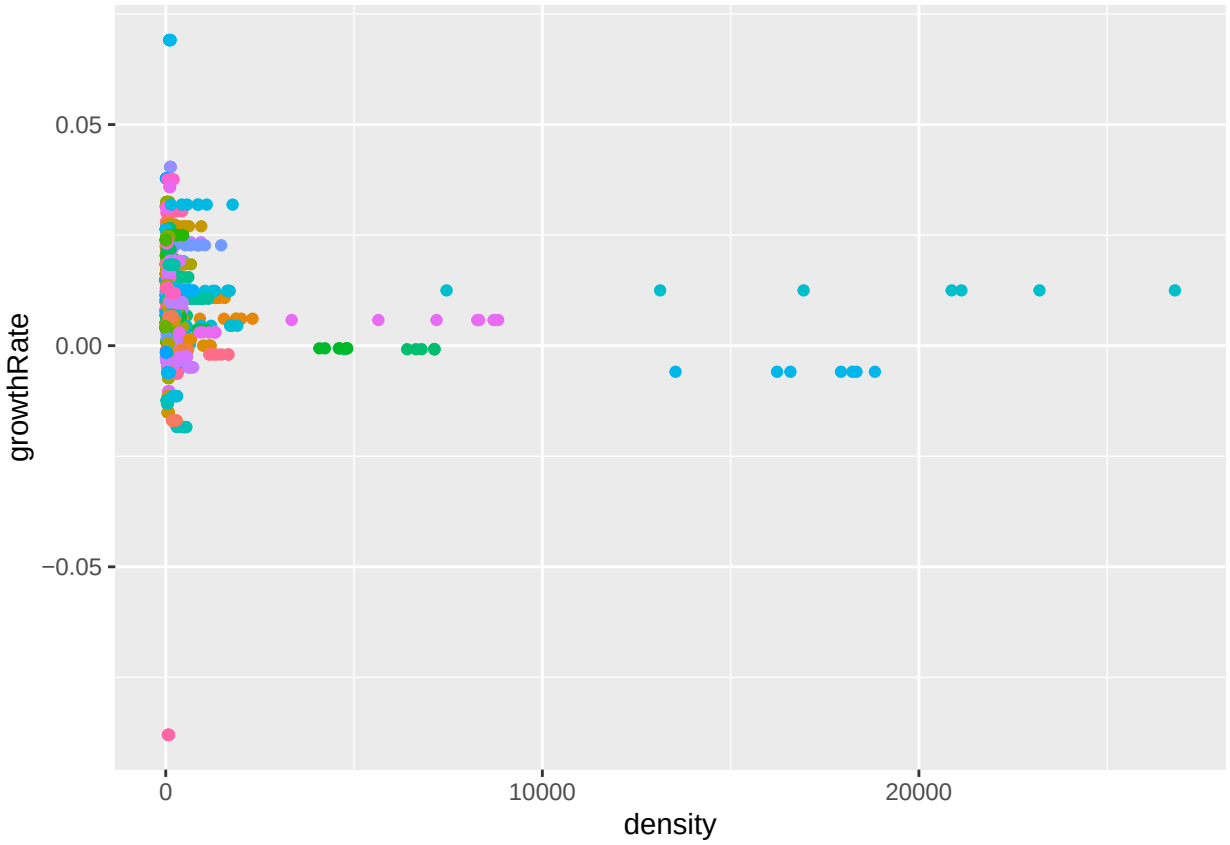
It looks like our big 5 nations are actually slowing down over time regarding their pop growth.

Now, I'm curious if density has any correlation with growth...

```r
pop_density <- mutate(pop, density = population/landAreaKm)
cor(pop_density$density, pop_density$growthRate)
```

```
## [1] -0.0761573
```

```r
ggplot(data = pop_density, mapping = aes(x = density, y = growthRate, color = country)) +
  geom_point(show.legend = FALSE)
```



It does appear that there is a decent enough negative correlation between population density and growth rate.