AQLI Test Packet - EPIC

2. Geospatial and Open-Source Section

Priyanka Gagneja

2023-11-24

## AQLI @ EPIC

library(magrittr)  
library(ggplot2)  
library(ggtext)  
library(patchwork)  
library(sf)  
  
library(gganimate)  
library(magick)  
library(camcorder)

gadm2\_aqli\_1998\_2021 <- read.csv("Inputs/task2/gadm2\_aqli\_1998\_2021.csv")

Lets look through the tasks and questions:

### 2.1 Basic wrangling

Q1. How many GADM2 regions are present in India?

GADM2\_region\_india\_count <- gadm2\_aqli\_1998\_2021 %>%   
 dplyr::filter(country == "India") %>%   
 dplyr::select(objectid\_gadm2) %>%   
 dplyr::distinct() %>%   
 dplyr::count()

A1. There are 690 GADM2 regions are present in India.

Q2. Calculate population weighted pollution average of all years at country (GADM0) level. a. Save the country level file as a CSV.

# population weighted pollution average of all years at country (GADM0) level  
# Initial understanding of the ques  
# gadm2\_aqli\_1998\_2021\_wt\_avg <- gadm2\_aqli\_1998\_2021%>%   
# dplyr::select(-contains("llpp\_")) %>%   
# dplyr::rowwise() %>%  
# dplyr::mutate(avg\_pm\_all\_yrs = mean(dplyr::c\_across(pm1998:pm2021), na.rm = TRUE) ) %>%   
# dplyr::ungroup() %>%   
# dplyr::group\_by(country) %>%   
# dplyr::mutate(total\_pop = sum(population, na.rm = TRUE),   
# wt\_avg\_pm = avg\_pm\_all\_yrs\*population/total\_pop\*100) %>%   
# dplyr::ungroup() %>%   
# dplyr::select(-contains("pm19"), -contains("pm20"))  
  
# NOTES for improvement  
# This section particularly runs slow. An alternate could be to use tidyr::pivot\_longer(), before averaging  
  
  
# Revised calculation of the ques  
# Assumption: country , name\_1, name\_2 depicts the country, state, city hierarchy  
# gadm2\_aqli\_1998\_2021\_wt\_avg <- gadm2\_aqli\_1998\_2021%>%  
# dplyr::select(-contains("llpp\_"), -whostandard, -natstandard, -iso\_alpha3) %>%  
# dplyr::group\_by(country) %>%  
# dplyr::mutate(total\_pop = sum(population, na.rm = TRUE),  
# wt = population/total\_pop\*100) %>%  
# dplyr::ungroup() %>%   
# dplyr::select(-population, -total\_pop) %>%  
# tidyr::pivot\_longer(cols = c(contains("pm")), values\_to = "pm\_values", names\_to = "year") %>%   
# dplyr::mutate(year = sub("pm", "", year),  
# pm\_wtd = wt\*pm\_values)  
  
# some data missingness experienced here for 7 countries. 1 came up due to presence of only 2 states and cities.  
# country s  
# <chr> <dbl>  
# 1 Akrotiri and Dhekelia 100  
# 2 Bouvet Island 0  
# 3 British Indian Ocean Territory 0  
# 4 Clipperton Island 0  
# 5 French Southern Territories 0  
# 6 Heard Island and McDonald Island 0  
# 7 South Georgia and the South Sand 0  
# 8 Svalbard and Jan Mayen 0  
  
# gadm2\_aqli\_country\_pop\_wt\_avg <- gadm2\_aqli\_1998\_2021%>%   
# dplyr::select(-contains("llpp\_")) %>%   
# dplyr::group\_by(country, year) %>%   
# dplyr::mutate(total\_pop = sum(population, na.rm = TRUE),   
# wt\_avg\_pm = avg\_pm\_all\_yrs\*population/total\_pop\*100) %>%   
# dplyr::ungroup() %>%   
# dplyr::select(-contains("pm19"), -contains("pm20"))  
#

# Final calculation of the ques  
# Assumption: country , name\_1, name\_2 depicts the country, state, city hierarchy  
gadm2\_aqli\_1998\_2021\_country\_avg\_pm <- gadm2\_aqli\_1998\_2021 %>%  
 dplyr::group\_by(country) %>%  
 dplyr::summarise(population = sum(population, na.rm = TRUE),  
 across(contains("pm"), ~mean(., na.rm = T))) %>%  
 dplyr::ungroup()   
# save the file  
write.csv(gadm2\_aqli\_1998\_2021\_country\_avg\_pm, "Outputs/task2/gadm2\_aqli\_1998\_2021\_country\_avg\_pm.csv")

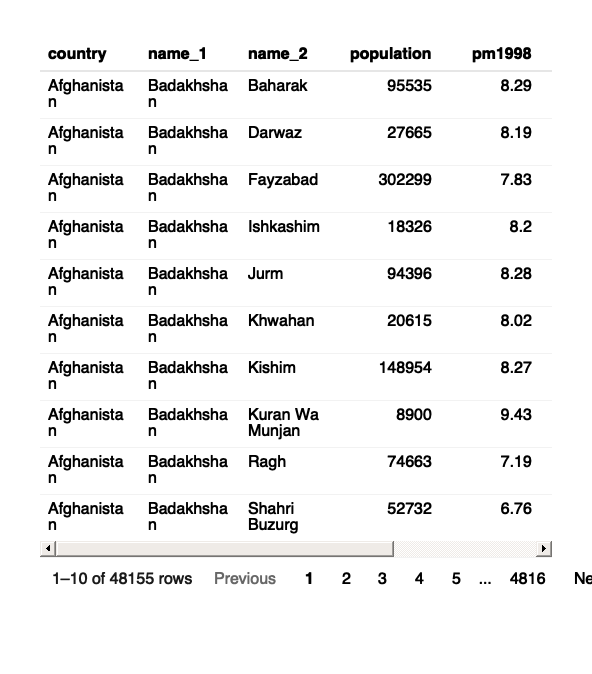
1. What are the 10 most polluted countries in 2021?

gadm2\_aqli\_1998\_2021\_country\_avg\_pm %>%   
 dplyr::arrange(-pm2021) %>%   
 dplyr::slice\_head(n = 10) %>%   
 gt::gt()

| country | population | pm1998 | pm1999 | pm2000 | pm2001 | pm2002 | pm2003 | pm2004 | pm2005 | pm2006 | pm2007 | pm2008 | pm2009 | pm2010 | pm2011 | pm2012 | pm2013 | pm2014 | pm2015 | pm2016 | pm2017 | pm2018 | pm2019 | pm2020 | pm2021 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Bangladesh | 164804810 | 44.08313 | 44.00031 | 49.81781 | 51.91156 | 51.77969 | 53.25609 | 52.46563 | 57.31188 | 62.16875 | 60.34219 | 63.33109 | 68.41281 | 63.76469 | 67.93031 | 68.82984 | 64.22453 | 66.99703 | 63.20516 | 64.09109 | 61.59562 | 73.59328 | 65.04031 | 73.51500 | 71.80078 |
| India | 1338260837 | 30.33465 | 30.50271 | 32.86830 | 36.50132 | 36.52790 | 39.42207 | 39.04874 | 39.90643 | 41.68568 | 43.12786 | 45.36106 | 45.33429 | 45.39801 | 48.49448 | 44.57386 | 46.25872 | 45.30070 | 45.23758 | 47.58012 | 45.98133 | 46.55990 | 48.22264 | 48.55128 | 51.09183 |
| Nepal | 30457421 | 24.44805 | 24.53519 | 32.27649 | 36.74532 | 33.10714 | 34.08688 | 35.44818 | 33.70442 | 38.30182 | 38.74675 | 39.40818 | 42.13078 | 35.68636 | 39.67104 | 38.37351 | 36.83792 | 37.44013 | 39.64974 | 40.89740 | 37.26351 | 40.56156 | 38.00234 | 38.39519 | 42.20442 |
| Pakistan | 238039784 | 24.33162 | 23.78144 | 25.47875 | 27.18862 | 26.62031 | 27.41475 | 27.21919 | 29.15638 | 32.17500 | 31.57250 | 30.44312 | 30.03981 | 33.30562 | 33.24544 | 31.43688 | 32.09412 | 31.72894 | 35.39744 | 35.12769 | 32.46550 | 30.41181 | 34.08481 | 34.94794 | 34.55494 |
| Myanmar | 55897221 | 26.34113 | 27.51187 | 24.53200 | 26.85800 | 27.24713 | 28.07562 | 31.75725 | 30.64800 | 30.91400 | 33.29725 | 30.36825 | 34.68012 | 32.89175 | 30.71862 | 31.59075 | 32.41437 | 33.62500 | 31.09150 | 31.42300 | 29.43150 | 31.10912 | 32.18800 | 32.84600 | 34.55325 |
| Democratic Republic of the Congo | 104994937 | 34.92356 | 26.53092 | 32.70423 | 35.01757 | 35.28381 | 33.49167 | 35.77280 | 35.48887 | 30.28151 | 33.69481 | 34.86849 | 34.60900 | 37.80159 | 36.00180 | 37.85586 | 39.42264 | 34.48782 | 36.90126 | 34.39025 | 30.38234 | 31.67377 | 34.72067 | 33.67201 | 33.83109 |
| Cameroon | 28568738 | 34.43810 | 32.10397 | 33.57879 | 35.50241 | 33.20052 | 33.26103 | 34.20621 | 34.19103 | 31.45534 | 32.51138 | 35.37655 | 33.35793 | 34.22293 | 33.10914 | 32.54241 | 33.69776 | 31.39517 | 30.29983 | 33.07948 | 33.83241 | 33.28603 | 32.18207 | 33.96259 | 33.57983 |
| Republic of the Congo | 5459096 | 28.64542 | 23.74771 | 28.29125 | 30.62458 | 30.03458 | 27.81563 | 32.07187 | 29.09563 | 25.63500 | 29.72792 | 33.93646 | 32.34333 | 34.12625 | 32.65271 | 32.28937 | 35.50396 | 32.41521 | 32.35063 | 32.84063 | 28.96500 | 32.02000 | 31.08438 | 33.08750 | 32.61792 |
| Rwanda | 12934347 | 29.11000 | 21.86100 | 33.36600 | 38.17100 | 34.40600 | 27.95367 | 35.66933 | 35.30133 | 30.36133 | 39.52667 | 37.81600 | 34.76000 | 36.60233 | 38.40067 | 37.61367 | 38.39400 | 34.20800 | 35.81067 | 35.08333 | 30.75833 | 30.58567 | 34.96000 | 34.20467 | 32.37200 |
| Burundi | 12242276 | 27.68218 | 20.48414 | 31.57842 | 37.17060 | 33.27639 | 29.62195 | 35.25519 | 34.95398 | 29.87008 | 38.82699 | 36.78654 | 34.10857 | 36.27068 | 37.09624 | 36.88774 | 38.27256 | 33.80624 | 35.50519 | 34.06338 | 29.14143 | 29.37902 | 33.72053 | 33.32594 | 32.15504 |

Q3. What was the most polluted GADM2 region in the world in 1998, 2005 and 2021?

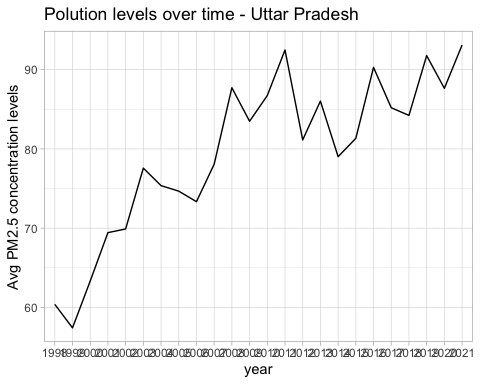
# Assuming Level 2 region in the question  
gadm2\_aqli\_1998\_2021 %>%   
 dplyr::select(country, name\_1, name\_2, population, pm1998, pm2005, pm2021) %>% # objectid\_gadm2,   
 reactable::reactable()



Most polluted cities or regions (with highest PM2.5 level concentrations).  
The reactable table can be sorted by clicking on the appropriate columns  
- 1998 - Unnao, UP, India  
- 2005 - NCT of Delhi, India  
- 2021 - NCT of Delhi, India

Q4. Plot a population weighted pollution average trendline plot for Uttar Pradesh from 1998 to 2021. Save this plot as a high-quality PNG file.

# slightly modified the plot for lack of clarity of ques on my part  
# gadm2\_aqli\_1998\_2021\_wt\_avg %>%   
# dplyr::filter(name\_1 == "Uttar Pradesh", name\_2 == "Agra") %>%   
# ggplot(aes(x = year, y = pm\_wtd)) +  
# geom\_line(group = 1) +   
# # facet\_wrap(~ name\_2)  
# labs(title = "Weighted PM levels over time") +   
# theme\_light()  
  
gadm2\_aqli\_1998\_2021 %>%  
 dplyr::group\_by(country, name\_1) %>%  
 dplyr::summarise(across(contains("pm"), ~mean(., na.rm = T))) %>%  
 dplyr::ungroup() %>%   
 dplyr::filter(name\_1 == "Uttar Pradesh") %>% #, name\_2 == "Agra"  
 tidyr::pivot\_longer(cols = c(contains("pm")), values\_to = "pm\_values", names\_to = "year") %>%   
 dplyr::mutate(year = sub("pm", "", year)) %>%   
 ggplot(aes(x = year, y = pm\_values)) +  
 geom\_line(group = 1) +   
 # facet\_wrap(~ name\_1) +  
 labs(title = "Polution levels over time - Uttar Pradesh",   
 y = "Avg PM2.5 concentration levels") +   
 theme\_light()

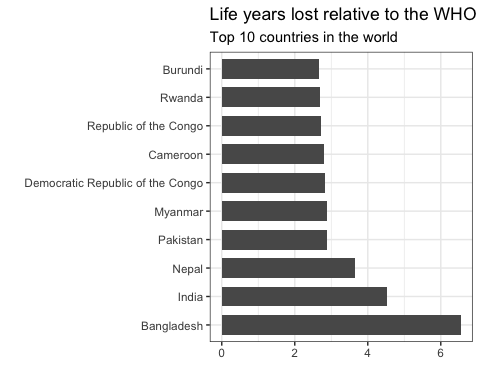


ggsave("Outputs/task2/plot\_04\_aqli.png")

### 2.2 Geospatial

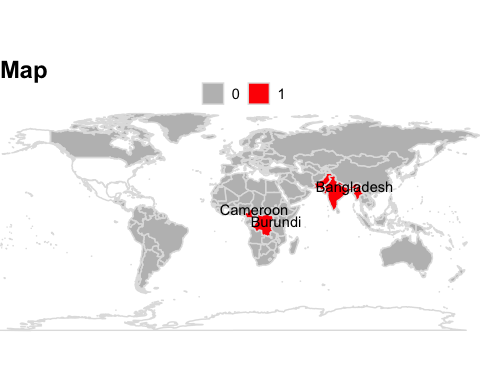
1. Plot a bar graph for the life years lost relative to the WHO guideline in the 10 most polluted countries in the world and plot them on a global country level map. For the map, the 10 most polluted country boundaries should be filled in with “dark red” and the rest of the map should be grayed out. Save both the bar graph and the map as high-quality PNG files.

top10\_polluted\_countries <- gadm2\_aqli\_1998\_2021\_country\_avg\_pm %>%   
 dplyr::arrange(-pm2021) %>%   
 dplyr::slice\_head(n = 10) %>%   
 dplyr::mutate(country = factor(country, levels = country))  
   
gadm2\_aqli\_1998\_2021\_country\_avg\_llpp <- gadm2\_aqli\_1998\_2021 %>%   
 dplyr::group\_by(country) %>%   
 dplyr::summarise(llpp\_who\_2021 = mean(llpp\_who\_2021, na.rm = T) ) %>%   
 dplyr::ungroup() %>%   
 dplyr::mutate(is\_top\_10 = dplyr::if\_else(country %in% top10\_polluted\_countries$country, "1", "0"))  
  
gadm2\_aqli\_1998\_2021\_country\_avg\_llpp %>%   
 dplyr::filter(country %in% top10\_polluted\_countries$country) %>%   
 ggplot(aes(x = factor(country, levels = levels(top10\_polluted\_countries$country)),   
 y = llpp\_who\_2021)) +   
 geom\_col(width = 0.7) +   
 labs(title = "Life years lost relative to the WHO guideline",  
 subtitle = "Top 10 countries in the world", y = "", x = "") +   
 coord\_flip() +  
 theme\_bw()



# library(tmap)  
# map\_obj + # tm\_shape(gadm2\_aqli\_1998\_2021\_country) +   
# tm\_polygons() +   
# tm\_shape(urban\_agglomerations) +   
# tm\_symbols(size = "population\_millions") +  
# tm\_facets(by = "year", nrow = 1, ncol = 1, free.coords = FALSE)  
  
  
# World map  
world <- map\_data("world") %>%  
 fortify()  
  
# Fonts  
f1 <- "Outfit"  
f2 <- "SF Mono"  
  
cols <- c("0" = "grey", "1" = "red")  
  
llpp\_map\_df <- world %>%   
 dplyr::left\_join(gadm2\_aqli\_1998\_2021\_country\_avg\_llpp, by = c("region" = "country"))  
  
top5\_labels <- top10\_polluted\_countries %>%   
 dplyr::slice\_head(n = 5) %>%   
 dplyr::inner\_join(llpp\_map\_df, by = c("country" = "region")) %>%   
 dplyr::group\_by(country) %>%   
 dplyr::summarise(long = mean(long, na.rm = T),  
 lat = mean(lat , na.rm = T)) %>%   
 dplyr::ungroup()  
  
top10\_labels <- top10\_polluted\_countries %>%   
 dplyr::inner\_join(llpp\_map\_df, by = c("country" = "region")) %>%   
 dplyr::group\_by(country) %>%   
 dplyr::summarise(long = mean(long, na.rm = T),  
 lat = mean(lat , na.rm = T)) %>%   
 dplyr::ungroup()

llpp\_map\_df %>%   
 ggplot() + # aes(fill = is\_top\_10)  
 geom\_polygon(aes(long, lat, group = group, fill = is\_top\_10), color="grey88") +   
 scale\_fill\_manual(values = cols, na.translate = FALSE) +  
 geom\_text(data = top10\_labels, nudge\_x = 3, nudge\_y = 3,  
 check\_overlap = TRUE,   
 aes(label = country, x = long, y = lat, group = country)) +   
 coord\_fixed(expand = FALSE) +  
 labs(title = "Map") +  
 theme\_void() + #base\_family = f1  
 theme(  
 legend.position = "top",  
 legend.title = element\_blank(),  
 legend.text = element\_text(size = 11),  
 plot.background = element\_rect(fill = "white", color = NA),  
 plot.title = element\_text(face = "bold", size = 18),  
 plot.subtitle = element\_text(size = 14)  
 )



1. Create a potential gain in life expectancy (relative to the WHO guideline) map of eastern v/s western europe at GADM level 2 and save it as a high-quality PDF.
2. Plot should be in AQLI “Potential gain in life expectancy” color scale. Visit AQLI website Index page > See legend for “Potential gain in life expectancy” and infer “exact” colors from that.
3. You can define east and west europe based on any acceptable definition online, but whatever definition you use - mention the source.
4. Feel free to add annotations/text boxes etc. to help explain the visualization.

##

1. Look at the AQLI website > switch to Air pollution tab > plot a static version of the global pollution map you see there, in those “exact” same colors. Export it as a high quality (320 dpi) SVG file.

# sf\_obj <- sf::st\_read(dsn = "Inputs/task2/gadm2\_aqli\_shapefile/aqli\_gadm2\_final\_june302023.shp",   
# quiet = TRUE) %>%   
# dplyr::rename(objectid\_gadm2 = obidgadm2) # country = name0,   
#   
# # gadm2\_aqli\_1998\_2021  
# map\_obj <- sf\_obj %>%   
# dplyr::select(objectid\_gadm2, geometry) %>%   
# dplyr::inner\_join(gadm2\_aqli\_1998\_2021, by = "objectid\_gadm2") %>%   
# dplyr::select(country, name\_1, name\_2, pm2021, geometry)  
#   
# map\_obj %>%   
# ggplot() +  
# geom\_sf(data = map\_obj, aes(fill = pm2021))