Assignment Intro HDS - Infant Death Scotland

Sophie Charlotte Zeiz

2024-11-18

```
library(readr)
library(dplyr)
library(here)
```

Infant mortality in Scotland

For this assignment, I wanted to look into the data from the **Scottish Public Health Observatory** regarding changes and geographical patterns in infant mortality. "The infant mortality rate is defined as the number of children who die before reaching their first birthday in a given year, expressed per 1 000 live births." (OECD/World Health Organization (2018)). Ending preventable deaths of newborns and children under the age of five is one of the World Health Organisations sustainable development goals 1. My target audience are health officers working in the NHS or the councils, as they could initiate further research or data acquisition. The dataset used for this assignment was extracted from the ScotPHO Online Profiles Tool and is freely available for the public. For additional analysis on geographical patterns, two datasets from **Public Health Scotland** were added: 1. the Urban-Rural-Classification was used to display differences in mortality rate between urban an rural areas in six categories (https://www.opendata.nhs.scot/dataset/urban-rural-classification) 2. to link both datasets, a codes and labels dataset from (https://opendata.scot/datasets/public+health+scotland-geography+codes+and+labels/)/

```
#Load data from ScotPHO

library(readr)
infant_death_all_geo <- read_csv("ScotPHO_datatab_infantdeath_update.csv")
glimpse(infant_death_all_geo)
```

```
## Rows: 752
## Columns: 11
## $ area_code                 <chr> "S00000001", "S00000001", "S00000001", "S000~
## $ area_type                 <chr> "Scotland", "Scotland", "Scotland", "Scotlan~
## $ area_name                 <chr> "Scotland", "Scotland", "Scotland", "Scotlan~
## $ year                      <dbl> 2004, 2005, 2006, 2007, 2008, 2009, 2010, 20~
## $ period                    <chr> "2002 to 2006 calendar years; 5-year aggrega~
## $ type_definition           <chr> "Crude rate per 1,000 live births", "Crude r~
## $ indicator                 <chr> "Infant deaths, aged 0-1 years", "Infant dea~
## $ numerator                 <dbl> 262.4, 262.6, 260.2, 253.8, 241.2, 240.2, 22~
## $ measure                   <dbl> 4.9, 4.8, 4.6, 4.4, 4.1, 4.1, 3.9, 3.7, 3.6,~
## $ upper_confidence_interval <dbl> 5.5, 5.4, 5.2, 5.0, 4.7, 4.6, 4.4, 4.2, 4.2,~
## $ lower_confidence_interval <dbl> 4.3, 4.2, 4.1, 3.9, 3.6, 3.6, 3.4, 3.2, 3.2,~
```

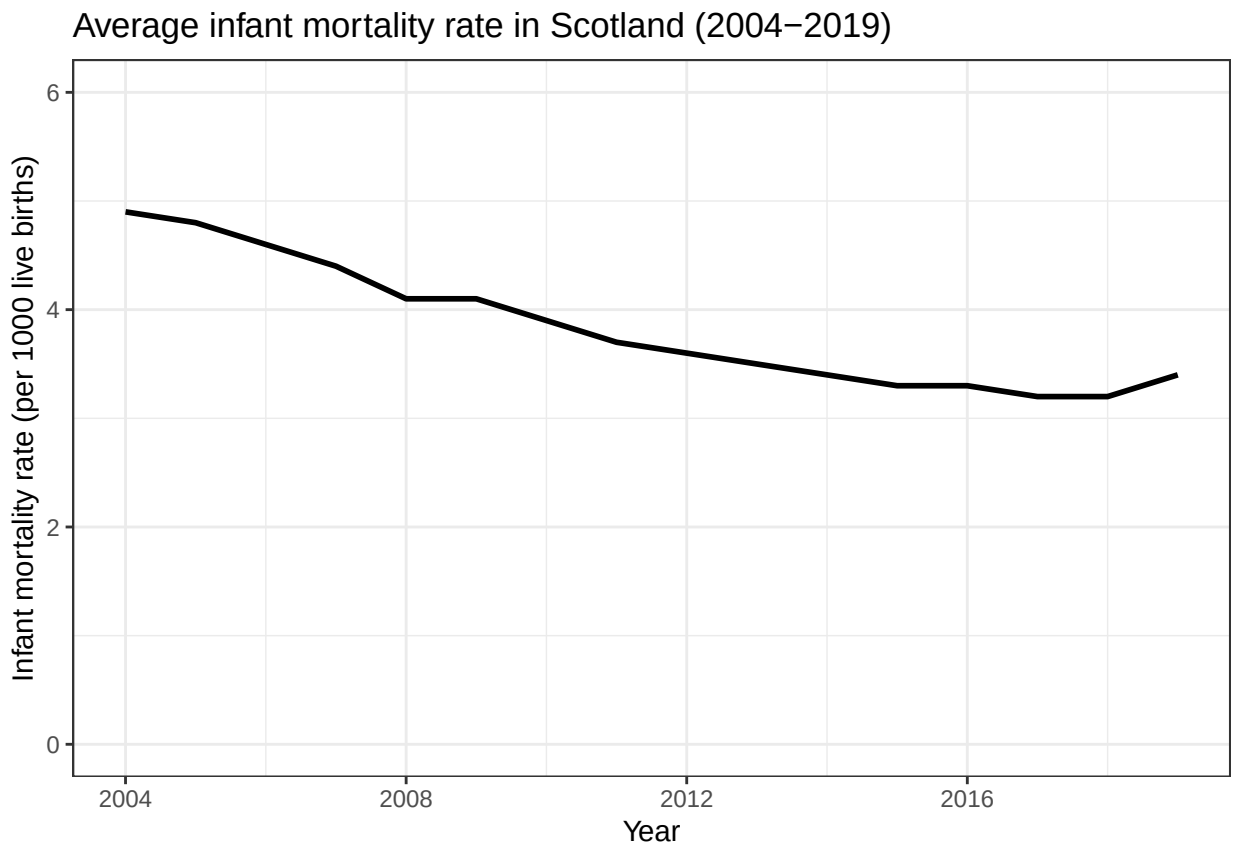
```r
#Clean data from ScotPHO

mortality_allgeo <- infant_death_all_geo %>%
  select(`area_code`,
         `area_type`,
         `area_name`,
         `year`,
         `measure`) %>%
  rename(mortality_rate = `measure`)
glimpse(mortality_allgeo)
```

```
## Rows: 752
## Columns: 5
## $ area_code      <chr> "S00000001", "S00000001", "S00000001", "S00000001", "S0~
## $ area_type      <chr> "Scotland", "Scotland", "Scotland", "Scotland", "Scotla~
## $ area_name      <chr> "Scotland", "Scotland", "Scotland", "Scotland", "Scotla~
## $ year           <dbl> 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2~
## $ mortality_rate <dbl> 4.9, 4.8, 4.6, 4.4, 4.1, 4.1, 3.9, 3.7, 3.6, 3.5, 3.4, ~
```

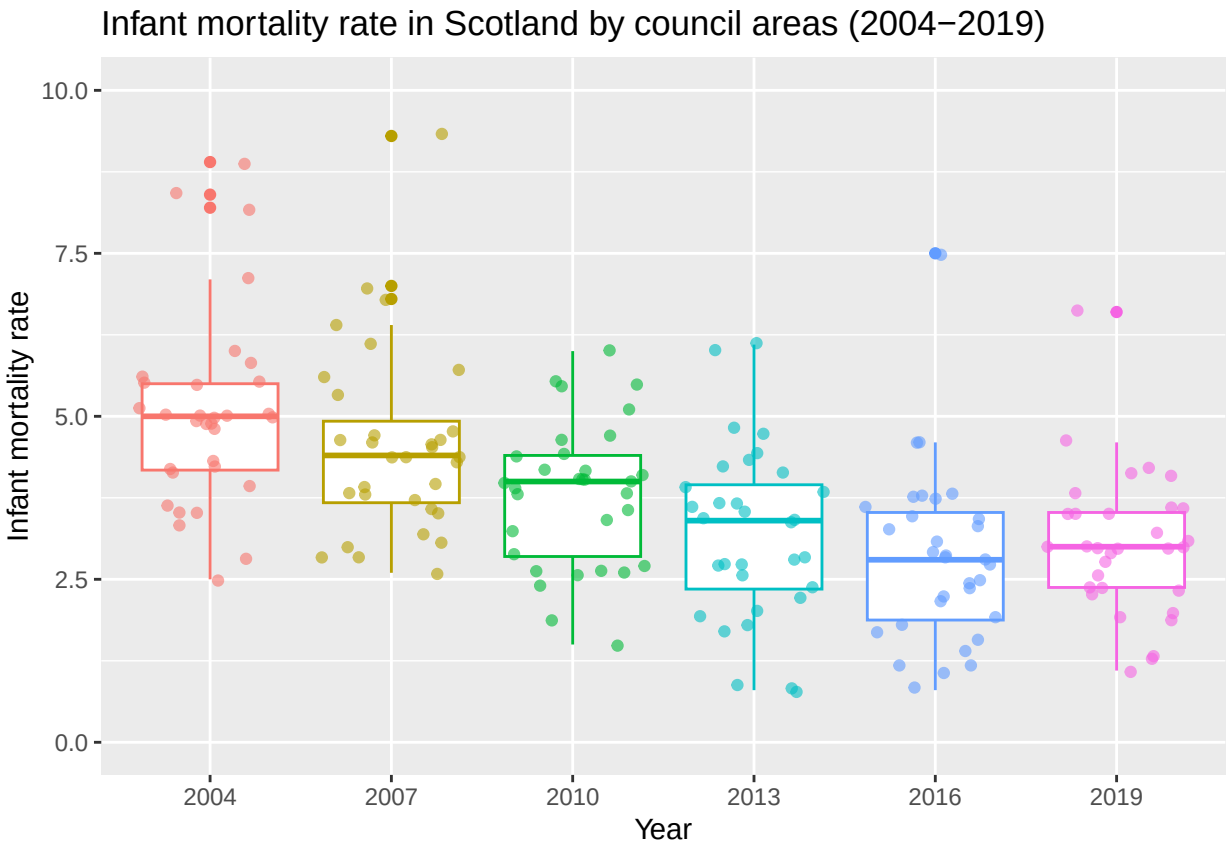
Infant mortality rate in Scotland between 2004-2019

```r
#visualize graph
library(ggplot2)
mortality_allgeo %>%
  select(year, mortality_rate, area_type) %>%
  filter(area_type == "Scotland") %>%
  ggplot() +
  geom_line(aes(x = year, y = mortality_rate), size = 1)+
  labs(x = "Year",
       y = "Infant mortality rate (per 1000 live births)",
       title = "Average infant mortality rate in Scotland (2004-2019)") +
  theme_bw()+
  scale_y_continuous(limits = c(0, 6)) #adjusting the scale to
```

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```




```
#data for boxplot
mortality_scot_by_council <- mortality_allgeo %>%
  select(`area_type`,
         `area_name`,
         `year`,
         `mortality_rate`) %>%
  filter(`area_type` == "Council area",
         `year` %in% c("2004", "2007", "2010","2013", "2016", "2019")) #every 3 years

mortality_scot_by_council$year <- as.factor(mortality_scot_by_council$year) #to help separate boxplots

#boxplot
library(ggplot2)

mortality_scot_by_council %>%
  ggplot(aes(x = factor(year),
             y = mortality_rate,
             colour = year)) +
  geom_boxplot() +
  geom_jitter(alpha = 0.6) +
  theme(legend.position = "none")+
  labs(x = "Year",
       y = "Infant mortality rate",
       title = "Infant mortality rate in Scotland by council areas (2004-2019)")+
  scale_y_continuous(limits = c(0, 10))
```



##Linking Urban-Rural classification

```
library(readr)
rural_class_complete <- read_csv("scottish-government-urban-rural-classification-2020-data-zone-2011-lo
View(rural_class_complete)
```

```
rural_6fold <- rural_class_complete %>%
  select(`DataZone`, `UR6FOLD`) %>%
  rename(area_code = `DataZone`,
         classification = `UR6FOLD`
         )
```

```
#read area dataset
library(readr)
area_to_hsc <- read_csv("dz2011_codes_and_labels_21042020.csv")


# connecting area code and hsc for 6 fold
CA_rural6fold <- rural_6fold %>%
  inner_join(`area_to_hsc`, by = c("area_code" = "DataZone")) %>%
  select(area_code, classification, CA, HSCP, HB)
```

```
library(dplyr)
get_mode <- function(x) {
  uniq_x <- unique(x)
```

```
  uniq_x[which.max(tabulate(match(x, uniq_x)))]}

# Combine council area, mortality rates and rural-urban classification

combined_mortalityallgeo_rural6fold <- mortality_allgeo %>%
  inner_join(`CA_rural6fold`, by = c("area_code" = "CA")) %>%
  select(area_type, area_name, year, mortality_rate, classification)
```

```
## Warning in inner_join(., CA_rural6fold, by = c(area_code = "CA")): Detected an unexpected many-to-ma
## i Row 241 of ‘x‘ matches multiple rows in ‘y‘.
## i Row 904 of ‘y‘ matches multiple rows in ‘x‘.
## i If a many-to-many relationship is expected, set ‘relationship =
##   "many-to-many"‘ to silence this warning.
```

```
#summarise mortality and classification per area_name and year
AVG_6_mort_class_percouncil_peryear <- combined_mortalityallgeo_rural6fold %>%
  group_by(area_name, year) %>%
  summarise(
    avg_mortality_rate = mean(mortality_rate, na.rm = TRUE),
    mode_classification = get_mode(classification) #mode for whole numbers of original categories
  )

#create classification categories via mode according to scot public health
 group_classification <- AVG_6_mort_class_percouncil_peryear %>%
  mutate(mode_classification = case_when(
    mode_classification == 1 ~ "Large Urban Areas",
    mode_classification == 2 ~ "Other Urban Areas",
    mode_classification == 3 ~ "Accessible Small Towns",
    mode_classification == 4 ~ "Remote Small Towns",
    mode_classification == 5 ~ "Accessible Rural",
    mode_classification == 6 ~ "Remote Rural",
    TRUE ~ "Unknown"
  ))

# average mortality rate per classification
avg_6class_mort_year <- AVG_6_mort_class_percouncil_peryear %>%
  select(`year`,`avg_mortality_rate`, `mode_classification`) %>%
  group_by(mode_classification, year) %>%
  summarise(AVG_mortality_rate = mean(avg_mortality_rate, na.rm = TRUE))

#r Perform Kruskal-Wallis test for 3 and 6 fold, include=FALSE}
kruskal_test_result <- kruskal.test(AVG_mortality_rate ~ mode_classification, data = avg_6class_mort_ye
kruskal_test_result
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  AVG_mortality_rate by mode_classification
## Kruskal-Wallis chi-squared = 13.619, df = 4, p-value = 0.008616
```

```
# -->continue with the 6-fold dataset
```

```r
avg_class_mort_year <- group_classification %>%
  select(`year`,`avg_mortality_rate`, `mode_classification`) %>%
  group_by(mode_classification, year) %>%
  summarise(avg_mortalityrate = mean(avg_mortality_rate, na.rm = TRUE))
```

```r
library(ggplot2)

#to keep order in the legend
avg_class_mort_year$mode_classification <- factor(avg_class_mort_year$mode_classification, levels = c("

avg_class_mort_year %>%
  select(`year`,`avg_mortalityrate`, `mode_classification`) %>%
  group_by(`mode_classification`) %>%
  ggplot(alpha= 1) +
  geom_line(aes(x= `year`, y= `avg_mortalityrate`, colour =`mode_classification`),size=1) +
              labs(x = "Year", y = "Infant mortality rate", title = "Infant mortality rate in Scotland
  theme(legend.position ="right")
```