Assignment Intro HDS - Infant Death Scotland

Sophie Charlotte Zeiz

2024-11-18

```r
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 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.

##Data aquisation The dataset used for this assignment was extracted from the ScotPHO Online Profiles Tool and is freely available for the public. For additional analysis data from the used the

```r
#load data from ScotPHO

library(readr)
infant_death_all_geo <- read_csv("ScotPHO_infant death all geo.csv")
glimpse(infant_death_all_geo)
```

```
## Rows: 1,248
## 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
mortality_allgeo <- infant_death_all_geo %>%
  select(`area_code`,
         `area_type`,
         `area_name`,
         `year`,
         `measure`) %>%
  rename(Year = `year`,
         Mortality_rate = `measure`)
head(mortality_allgeo)
```

```
## # A tibble: 6 x 5
##   area_code area_type area_name  Year Mortality_rate
##   <chr>     <chr>     <chr>     <dbl>          <dbl>
## 1 S00000001 Scotland  Scotland   2004            4.9
## 2 S00000001 Scotland  Scotland   2005            4.8
## 3 S00000001 Scotland  Scotland   2006            4.6
## 4 S00000001 Scotland  Scotland   2007            4.4
## 5 S00000001 Scotland  Scotland   2008            4.1
## 6 S00000001 Scotland  Scotland   2009            4.1
```

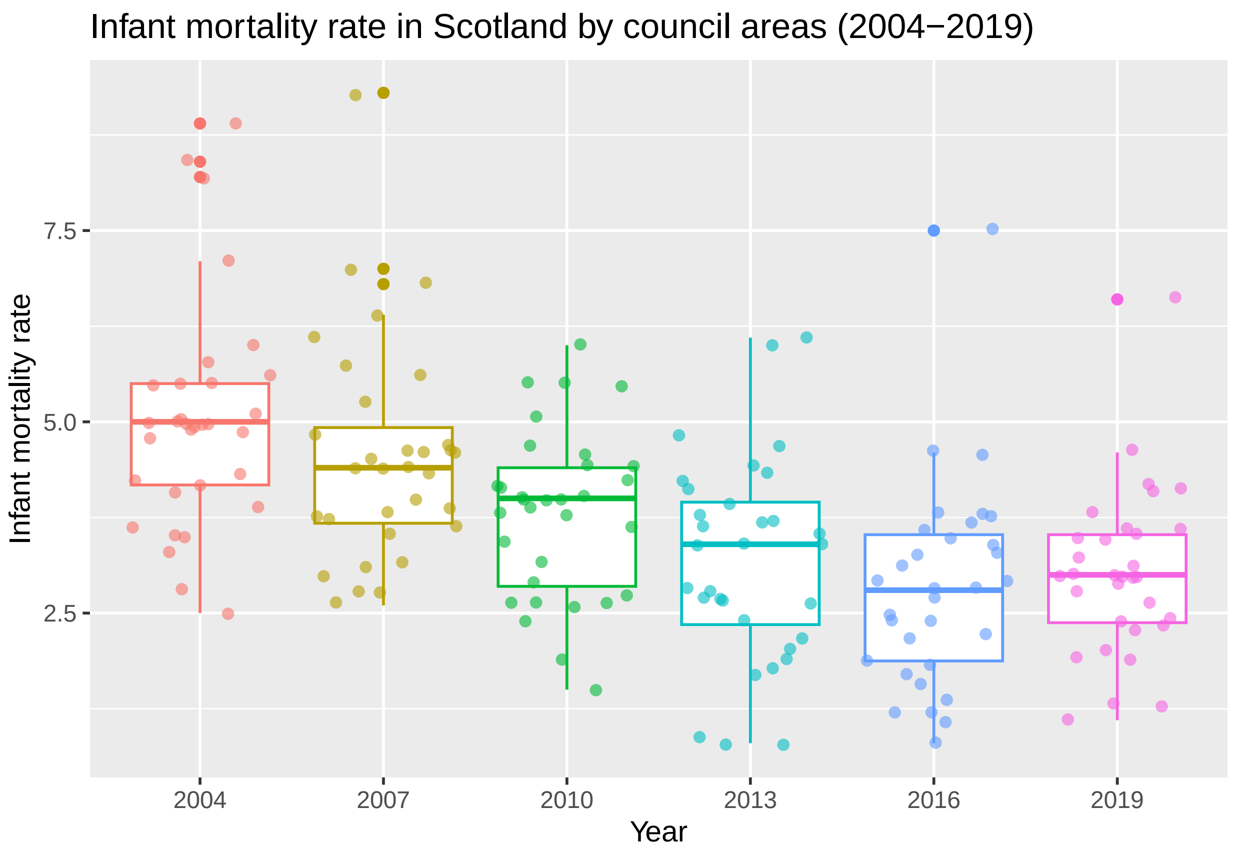
Infant mortality rate in Scotland between 2004-2019

```r
#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)")
```



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

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

```r
#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)
```

```r
library(dplyr)
get_mode <- function(x) {
  uniq_x <- unique(x)
  uniq_x[which.max(tabulate(match(x, uniq_x)))]
}
```

```r
# 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-man
## 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.
```

```r
#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 for 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_yea
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
```

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

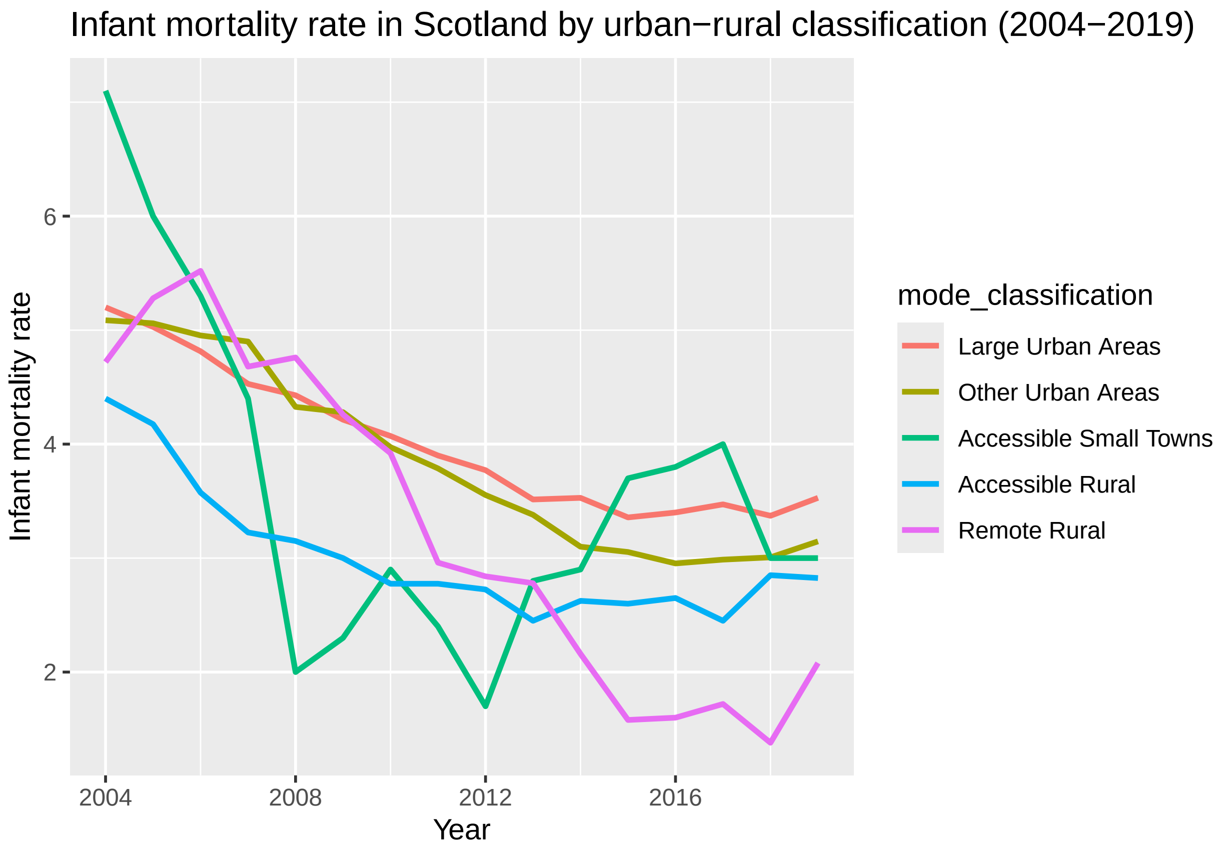
```r
avg_class_mort_year <- group_classification %>%
  select(`Year`,`avg_mortality_rate`, `mode_classification`) %>%
  group_by(mode_classification, Year) %>%
  summarise(avg_mortalityrate = mean(avg_mortality_rate, na.rm = TRUE))
```

```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")
```

```
## 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.
```


Infant mortality rate in Scotland by urban−rural classification (2004−2019)
Infant mortality rate
Year
mode_classification
Large Urban Areas
Other Urban Areas
Accessible Small Towns
Accessible Rural
Remote Rural