

Assignment Intro HDS - Infant Death Scotland

Sophie Charlotte Zeiz

Infant mortality in Scotland

Introduction

For this assignment, I wanted to look into the data from the **Scottish Public Health Observatory** regarding changes and distribution in infant mortality in Scotland. “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.” (1). Ending preventable deaths of newborns and children under the age of five is one of the World Health Organisations sustainable development goals. (2) My target audience are professionals working for the NHS or councils. The dataset used for this assignment was extracted from the ScotPHO *Online Profiles Tool* and is freely available for the public.

Questions

With my visualisations I want to answer the following Questions:

1. How did the infant mortality rate in Scotland change between 2004 and 2019?
2. How is the data distributed around the Scottish average, are there outliers?
3. How has the infant mortality rate of the outlying council(s) and their healthboard(s) developed?

Preparation

Load packages

```
#loading the packages i used most during the course
library(readr)
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
```

Read data from ScotPHO

The data was downloaded from the Scottish Public Health Observatory *Online Profiles Tool* (3). The dataset is called Infant Death, aged 0-1 years. The dataset used for this report was last updated in February of 2023.

```
#Read in data from ScotPHO
```

```
library(readr)
infant_death_all_geo <- read_csv("~/University of Aberdeen/Introduction to Health Data Science/Assignment
```

Inspect data

The dataset provides values for the years 2004 up to 2019. Geographical areas available are: Council area, Health board, HSC partnership, Scotland. The dataset includes raw numbers as well as the mortality rate with CIs.

```
glimpse(infant_death_all_geo) #to see what data dataset contains
```

```
## Rows: 1,248
## Columns: 11
## $ area_code      <chr> "S000000001", "S000000001", "S000000001", "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,~
```

```
any(is.na(infant_death_all_geo)) #to see if there are missing values
```

```
## [1] FALSE
```

Data cleaning

To be able to compare between areas with different population sizes, i chose to only work with the infant mortality rate (“measure”), not the total numbers (“numerator”). While observing the *mortality_allgeo* dataset, we can observe that every variable has its own column.

```
#Selecting values i need for my visualisations
```

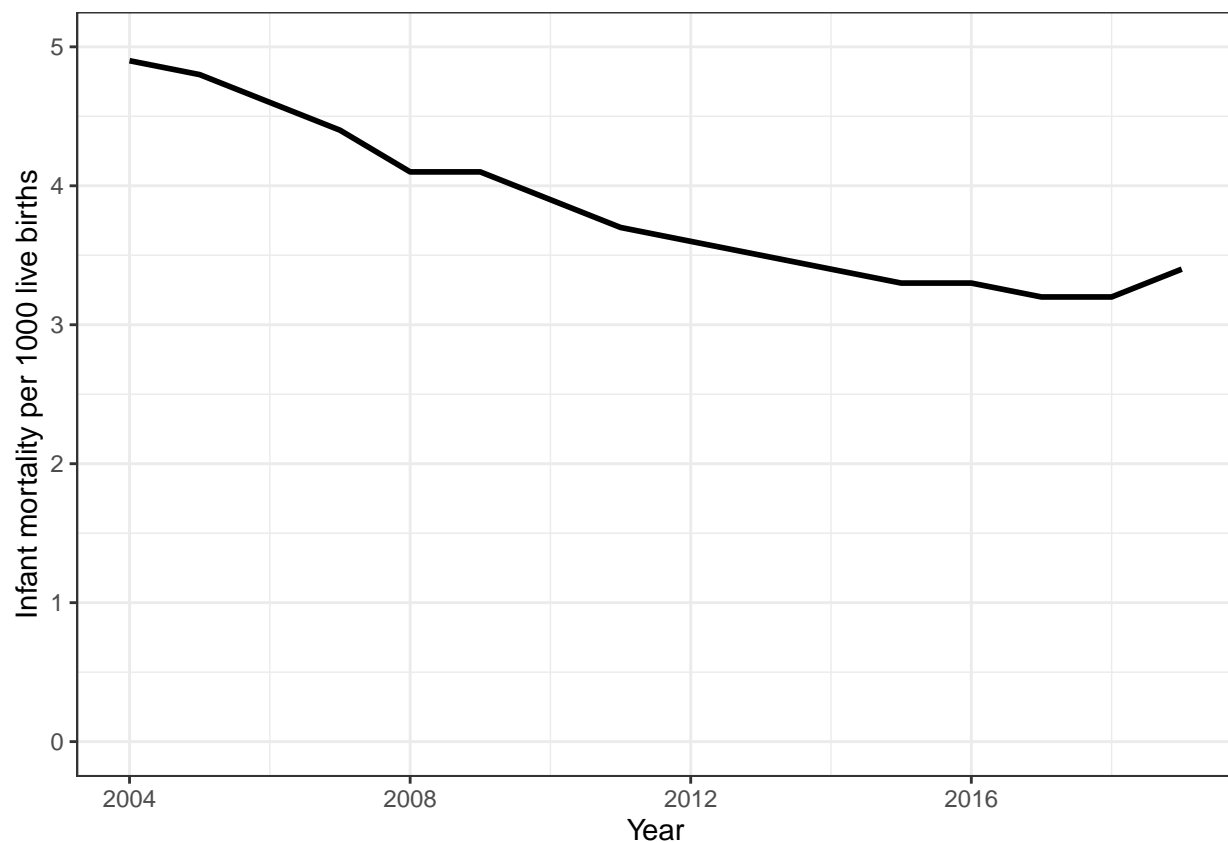
```
mortality_allgeo <- infant_death_all_geo %>%
  select(area_code,
         area_type,
         area_name,
         year,
         measure) %>%
  rename(mortality_rate = `measure`)
```

Visualisation

How did the infant mortality rate in Scotland change between 2004 and 2019?

This line graph shows the history of the infant mortality rate in Scotland. Between 2004 and 2018 we can observe a reduction in infant mortality, in 2019 the rate increased slightly.

```
#visualize graph 1
library(ggplot2)
mortality_allgeo %>%
  select(year, mortality_rate, area_type) %>%
  filter(area_type == "Scotland") %>%
  ggplot() +
  geom_line(aes(x = year, y = mortality_rate), linewidth = 1)+
  labs(x = "Year",
       y = "Infant mortality per 1000 live births") +
  theme_bw()+
  scale_y_continuous(limits = c(0, 5)) #adjusting the scale to include zero -> more realistic visualis
```



How is the data distributed around the Scottish average, are there outliers?

To display the reduction of infant mortality rate within Scotland, I chose to create boxplots using the data from the 33 council areas. ### Prepare data for boxplots

```

#data for boxplot
mortality_scot_by_council <- mortality_allgeo %>%
  select(`area_type`,
         `area_name`,
         `year`,
         `mortality_rate`) %>%
  filter(`area_type` == "Council area")

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

```

Is the data normally distributed or skewed? Since i wanted to include the SD or IQR I needed to know if the data is normal distributed or skewed. I used the Q-Q-plot and histogram. The histogram showed an outlier to the right, that's why I also used the Q-Q-Plot to make sure. In the *mortality_scot_by_council* dataset we have 192 observations, which approximately allows for 14 bins (square route of $192 = 13.86$).

```

#Histogram
mortality_scot_by_council %>%
  ggplot(aes(x = mortality_rate))+
  geom_histogram(bins = 14)

#Q-Q-plot
mortality_scot_by_council %>%
  ggplot(aes(sample = mortality_rate))+
  geom_qq()+
  geom_qq_line()

#data is skewed --> IQR instead of SD (i think, not 100% sure but had to decide)

```

Preparing additional data for boxplot I created a dataset only including to outliers in 2019 on the higher end since a much lower mortality rate isn't a big problem.

```

#IQR values per year
IQR_data <- mortality_scot_by_council %>%
  group_by(year) %>%
  summarise(
    Q1 = quantile(mortality_rate, 0.25, na.rm = TRUE),
    Q3 = quantile(mortality_rate, 0.75, na.rm = TRUE),
    IQR = Q3 - Q1,
    upper_whisker = Q3 + 1.5 * IQR)

# for upper outliers by year
mortality_outliers <- mortality_scot_by_council %>%
  select(area_name, year, mortality_rate) %>%
  left_join(IQR_data, by = "year") %>%
  select(area_name, year, mortality_rate, upper_whisker) %>%
  filter(mortality_rate > upper_whisker)

#mortality outliers for boxplot years
outliers_boxplot <- mortality_outliers %>%
  filter(year %in% c(2004, 2007, 2010, 2013, 2016, 2019))

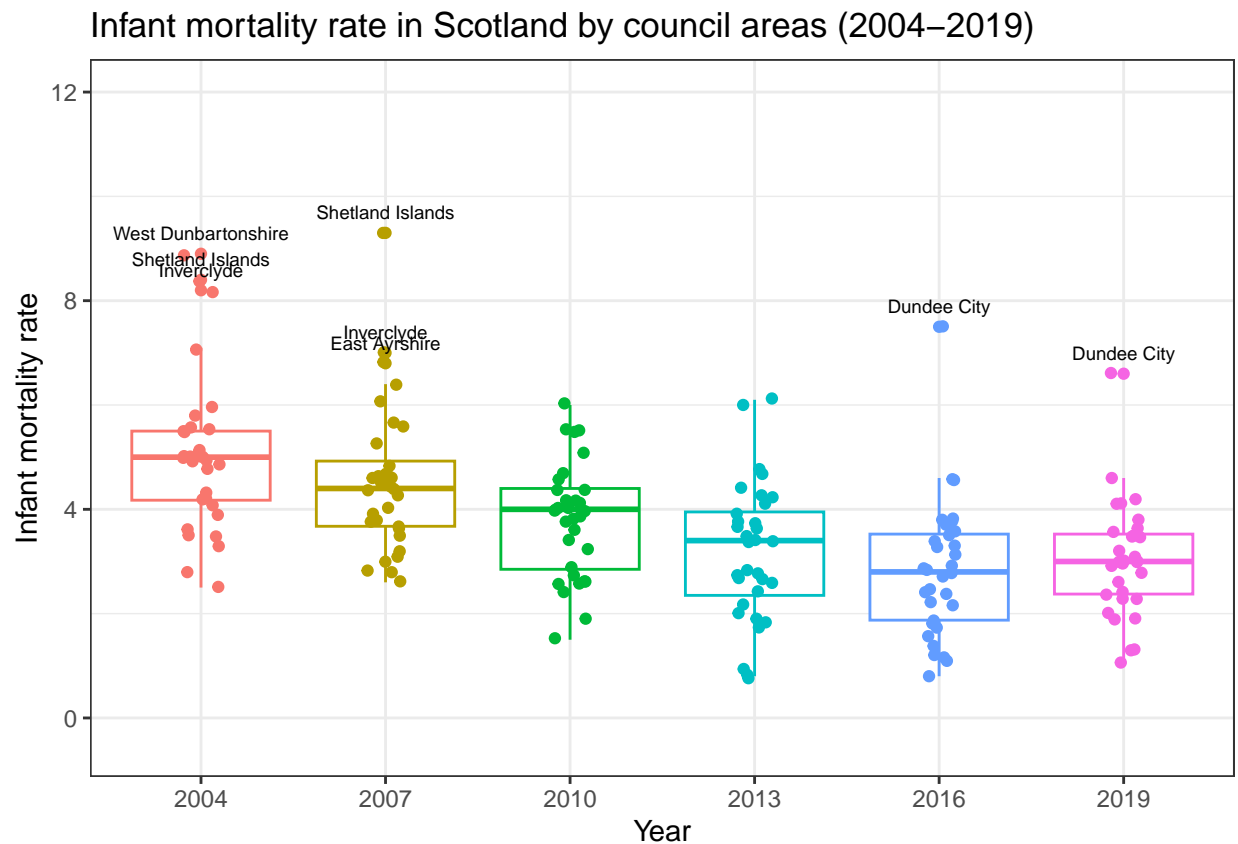
```

```

#boxplot
library(ggplot2)

mortality_scot_by_council %>%
  filter(year %in% c(2004, 2007, 2010, 2013, 2016, 2019)) %>%
  ggplot(aes(x = factor(year),
             y = mortality_rate,
             colour = year)) +
  geom_boxplot() +
  geom_jitter(alpha = 1, width = 0.1)+
  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(-0.5, 12))+
  geom_text(data= outliers_boxplot, aes(x = year, y = mortality_rate, label = area_name),
           color = "black",
           size = 2.5,
           vjust = -1,
           position = position_jitter(width = 0, height = -1))+ #adjustment of positions with help of
  theme_bw()+
  theme(legend.position = "none")

```



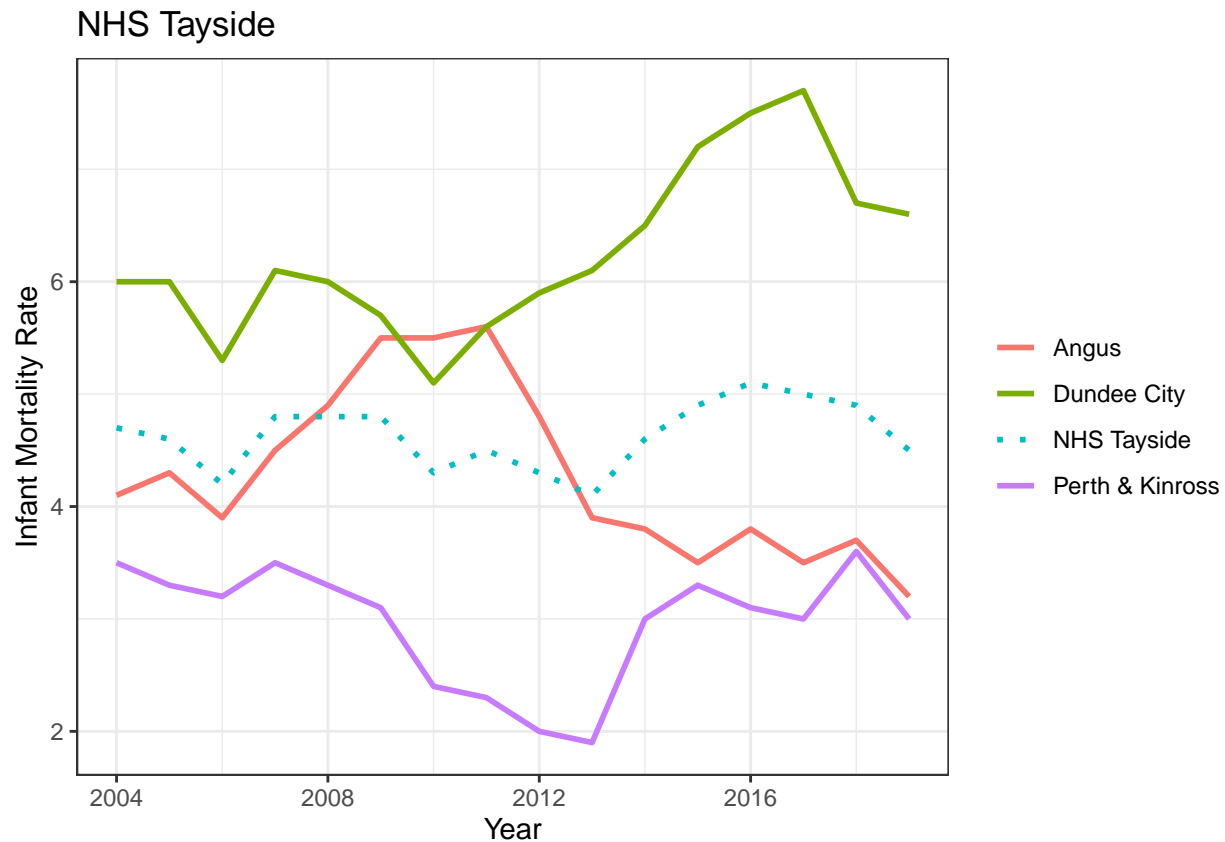
With an infant mortality rate of 6.6 in 2019, **Dundee City** was the only council higher than 1.5 IQR points of the third quartile.

How has the infant mortality rate of the outlying council(s) and their health-board(s) developed?

The **Dundee City** council forms the **NHS Tayside** together with the councils **Angus** and **Perth & Kinross**.

```
#linegraph for NHS Tayside (dotted) and the councils (solid)

library(ggplot2)
mortality_allgeo %>%
  select(area_name, year, mortality_rate) %>%
  filter(area_name %in% c("NHS Tayside", "Dundee City", "Perth & Kinross", "Angus")) %>%
  ggplot()+
  geom_line(aes(x = year, y = mortality_rate, colour = area_name, linetype = area_name), linewidth = 1)+
  scale_linetype_manual(values = c("NHS Tayside" = "dotted",
                                   "Dundee City" = "solid",
                                   "Perth & Kinross" = "solid",
                                   "Angus" = "solid")) +
  labs(x = "Year", y = "Infant Mortality Rate", title = "NHS Tayside")+
  theme_bw()+
  theme(legend.title = element_blank())
```



Summary

This report aims to provide an analysis of infant mortality trends in Scotland, targeting professionals who work for health boards and councils, especially the NHS Tayside and Dundee City council area. The key message of this report is to highlight the national and regional variations in infant mortality rates, offering insights into areas that may require further investigation to explain the observed differences or targeted interventions to reduce preventable deaths.

Data Used

The visualisations in this report are based on data from the Scottish Public Health Observatory (ScotPHO) Online Profiles Tool. The data used for these visualisations is the Infant Death, ages 0-1 dataset which was last updated in February 2023. The dataset covers yearly infant mortality rates, CI's and total infant death numbers from 2004 to 2019, between Scotland, NHS health boards, HSC partnership and council areas, providing a broad view of trends over time at the national and regional levels.

Data Limitations

Timeframe: The most recent dataset ends in 2019, meaning the data is nearly five years old. The latest trends and changes aren't taken into account.

Lack of Detail: The dataset does not specify the causes of death or the location (hospital or home), which are critical for understanding the underlying factors contributing to mortality rates.

Strengths and Limitations of the Visualisations

Line graph: National Trend of Infant Mortality

This graph shows the overall decline in national infant mortality rates, providing an accessible summary of Scotland's progress. The graph could benefit from further data distinguishing the numbers between preventable and non-preventable deaths, like death due to sudden infant death syndrome, which would help target interventions more effectively.

Boxplots: Regional spread and outliers The boxplots show the distribution of mortality rates across councils around the median and in the interquartile range. By clearly highlighting outliers, council areas that should examine their infant mortality rate closer, like Dundee City in 2019, can be suggested. The choice of three-year intervals prevents information overload, but may miss finer trends within each year.

Line Graph: Health Board-Level Analysis: NHS Tayside (Angus, Dundee City, Perth & Kinross)

This line graph provides a clear comparison of infant mortality trends within the Tayside health board, highlighting the elevated rates in Dundee City. Choosing a dotted line for the health board and solid lines for the councils aids readability.

References

(1): <https://www.oecd.org/en/data/indicators/infant-mortality-rates.html> (2): https://www.who.int/data/gho/data/themes/topics/sdg-target-3_2-newborn-and-child-mortality (3): https://scotland.shinyapps.io/ScotPHO_profiles_tool/