Report Introduction to Health Data Science

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) Looking into the data around infant death is important, because 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*. It is called **"Infant Death, aged 0-1 years"** and was last updated in February of 2023.(3)

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 health board(s) developed?

Preparation

```r
#Load packages
library(readr)
library(dplyr)
library(tinytex)
```

Read data from ScotPHO

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

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 CI's.

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

```
## 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,~
```

```
any(is.na(infant_death_all_geo)) #to check for 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 variables I need for my visualisations in general

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

Visualisation

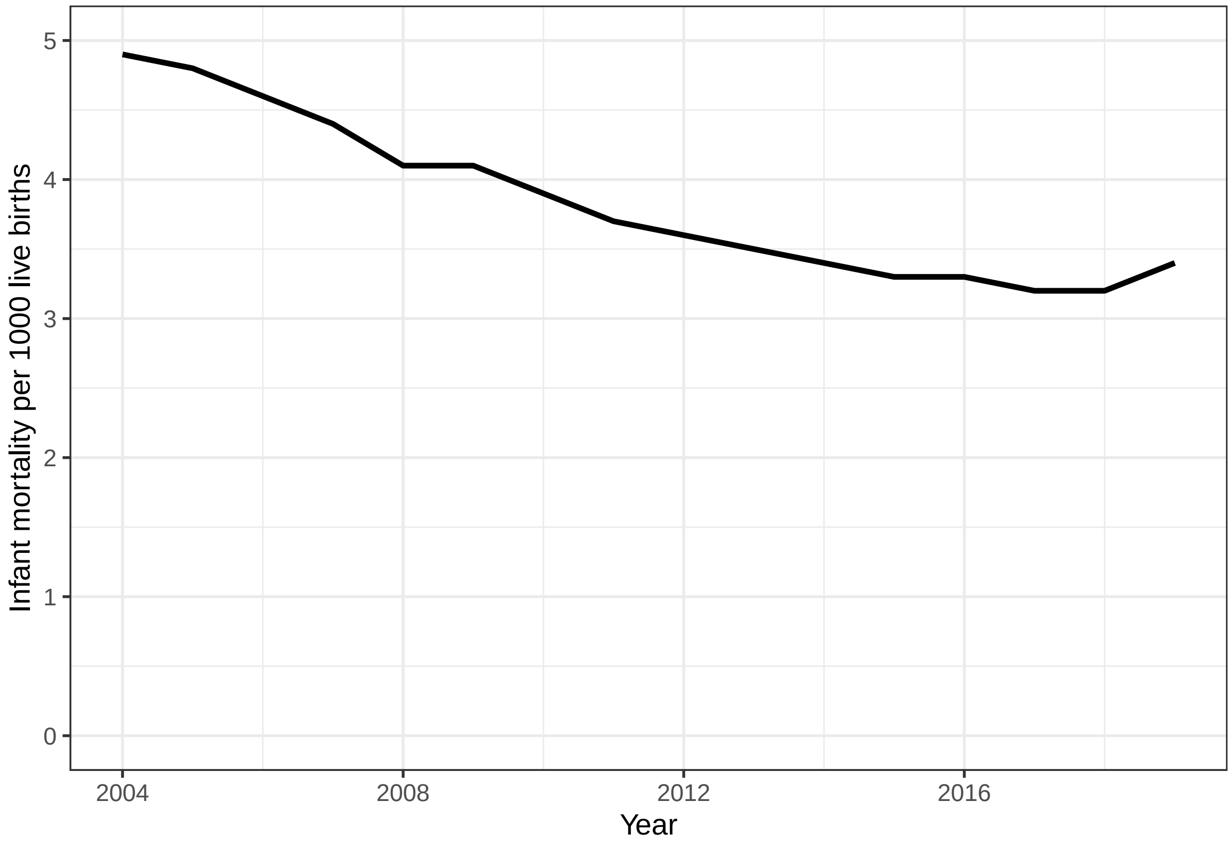
How did the infant mortality rate in Scotland change between 2004 and 2019?

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 to include zero -> more realistic visualisation (4)
```



Test and prepare data for graph 2

```
#prepare data
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 box plots per year in one graph (4)
```

```
#Since I wanted to include the SD or IQR I need to know if the data is normal distributed or skewed:
#Histogram
mortality_scot_by_council %>%
  ggplot(aes(x = `mortality_rate`))+
  geom_histogram(bins = 14) #(square root of 192 = 13.86 -> 14)

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

#data is skewed --> IQR + box plot instead of SD
```

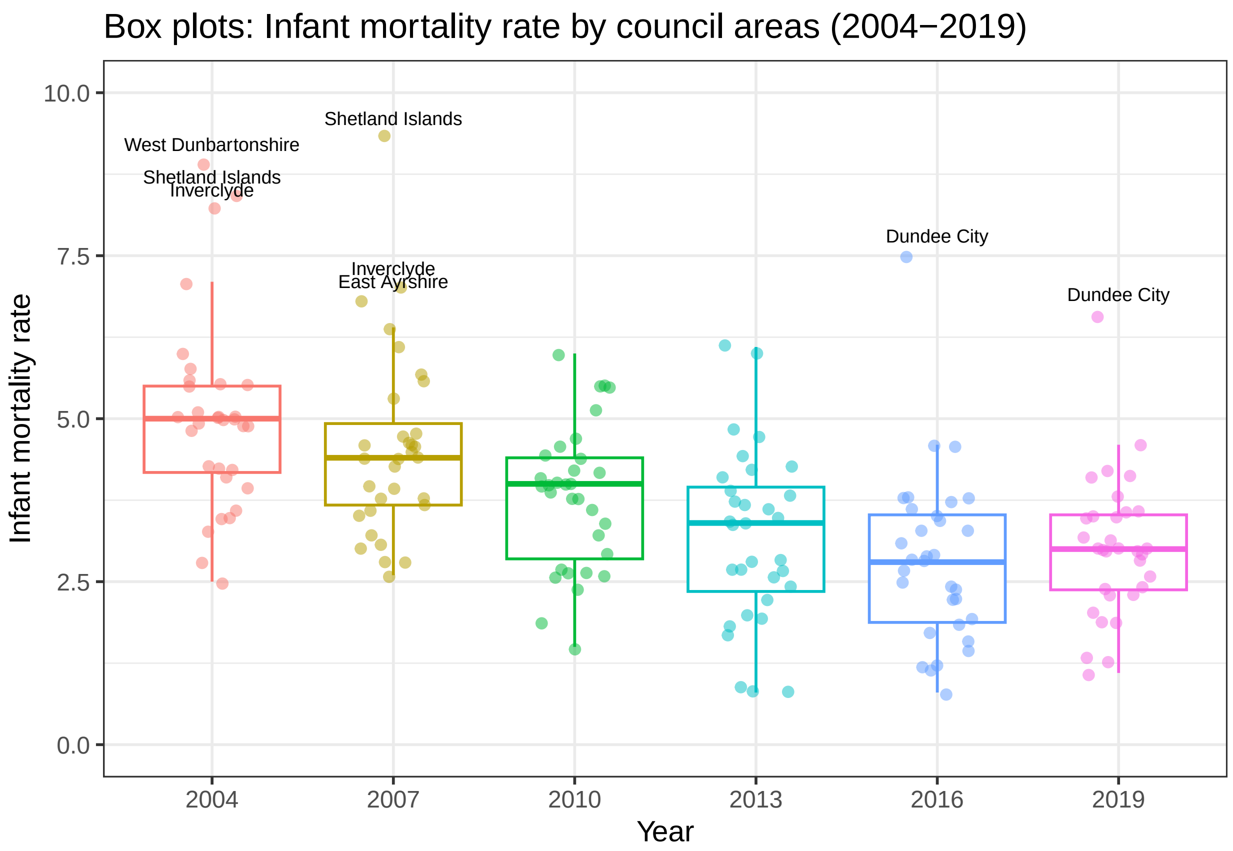
```
#IQR values per year
IQR_data <- mortality_scot_by_council %>%
group_by(`year`) %>%
summarise(
Q1 = quantile(`mortality_rate`, 0.25),
Q3 = quantile(`mortality_rate`, 0.75),
IQR = Q3 - Q1,
upper_whisker = Q3 + 1.5 * IQR)

# for upper outliers by year
mortality_outliers <- mortality_scot_by_council %>%
  select(`area_name`, `year`, `mortality_rate`) %>%
  left_join(IQR_data, by = "year") %>% #(4)
  select(`area_name`, `year`, `mortality_rate`, `upper_whisker`) %>%
  filter(`mortality_rate` > `upper_whisker`) #only including high outliers

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

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

```
library(ggplot2)
mortality_scot_by_council %>%
  filter(`year` %in% c(2004, 2007, 2010, 2013, 2016, 2019)) %>%
  ggplot(aes(x = factor(`year`), #(4)
             y = `mortality_rate`,
             colour = `year`)) +
  geom_boxplot(outlier.color=NA) + #(5) to remove the doubled outlier from jitter
  geom_jitter(alpha = 0.5, width = 0.2)+
  theme(legend.position = "none")+
  labs(x = "Year",
       y = "Infant mortality rate",
       title = "Box plots: Infant mortality rate by council areas (2004-2019)")+
  scale_y_continuous(limits = c(0, 10))+
  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 = -2, height = -0.2))+ #adjustment of positions: (4)
  theme_bw()+
  theme(legend.position = "none")
```



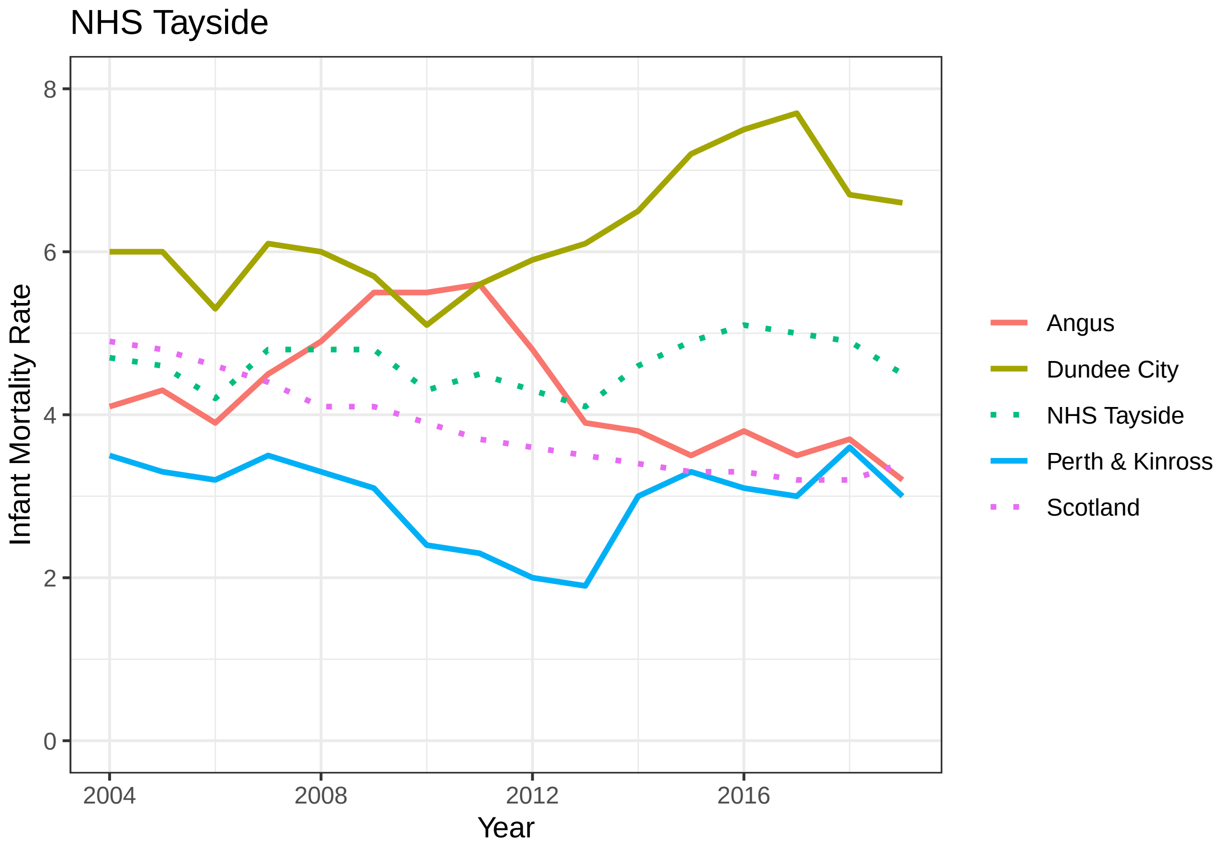
This boxplot displays the distribution of the infant mortality rate of the 33 council areas as well as outliers in a 3 year interval. In 2019, the council area of **Dundee City** was the only outlier (1.5 IQR points over the the third quartile) with an infant mortality rate of 6.6.

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**.

```r
#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", "Scotland")) %>%
  ggplot()+
  geom_line(aes(x = `year`, y = `mortality_rate`, colour = `area_name`, linetype = `area_name`), linewic
  scale_y_continuous(limits = c(0, 8))+
  scale_linetype_manual(values = c("NHS Tayside" = "dotted",
                                   "Dundee City" = "solid",
                                   "Perth & Kinross" = "solid",
                                   "Angus" = "solid",
                                   "Scotland" = "dotted")) +
  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, differentiating between Scotland, health boards, HSC partnerships 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 can't be 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

Graph 1 - line graph: National Trend of Infant Mortality

This graph shows the overall decline in infant mortality rate, 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.

Graph 2 - box plots: Regional spread and outliers The box plots show the distribution of mortality rates across councils around the median and the interquartile range. By 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.

Graph 3 - multiple lines 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. Adding the Scottish trend adds context and perspective to the graph. Choosing a dotted line for the health board and Scotland and solid lines for the councils aids readability.

References

1. Infant Mortality Rates. Organisation for Economic Co-operation and Development. Cited: 12 Oct 2024. Available from: https://www.oecd.org/en/data/indicators/infant-mortality-rates.html

2. SDG Target 3.2 End preventable deaths of newborns and children under 5 years of age. World Health Organisation. Cited: 12 Oct 2024. Available from: https://www.who.int/data/gho/data/themes/topics/sdg-target-3_2-newborn-and-child-mortality)

3. ScotPHO Online Profiles Tool - Infant Death, 0-1 year. Scottish Health Observatory. Last updated: Feb 2023. Downloaded: 18 Oct 2024 & 28 Nov 2024. Available from: https://scotland.shinyapps.io/ScotPHO_profiles_tool

4. ChatGTP. Available from: https://chatgpt.com/

5. extra point at boxplot with with jittered points (ggplot2). Stack Overflow. 2020. Cited: 04. Dec 2024. Available from: https://stackoverflow.com/questions/63784089/extra-point-at-boxplot-with-with-jittered-points-ggplot2

Use of GenAI

I acknowledge the use of ChatGPT from https://chatgpt.com/ to:
- improve readability of my text
- help me understand errors and warnings returned from RStudio
- troubleshoot my code in case I could not make it work (e.g. define years as a factor in a dataset to separate box plots by year)
- fine tuning plots in terms of layout (adjusting geom text in my box plot with "position = position_jitter")
I commented in the code where I used ChatGTP with the reference (4).