Data Visualization II: Multivariate Plots - Group Work

Radhika Ananth and Gaea Morales

Summer 2022

Multivariate Plots

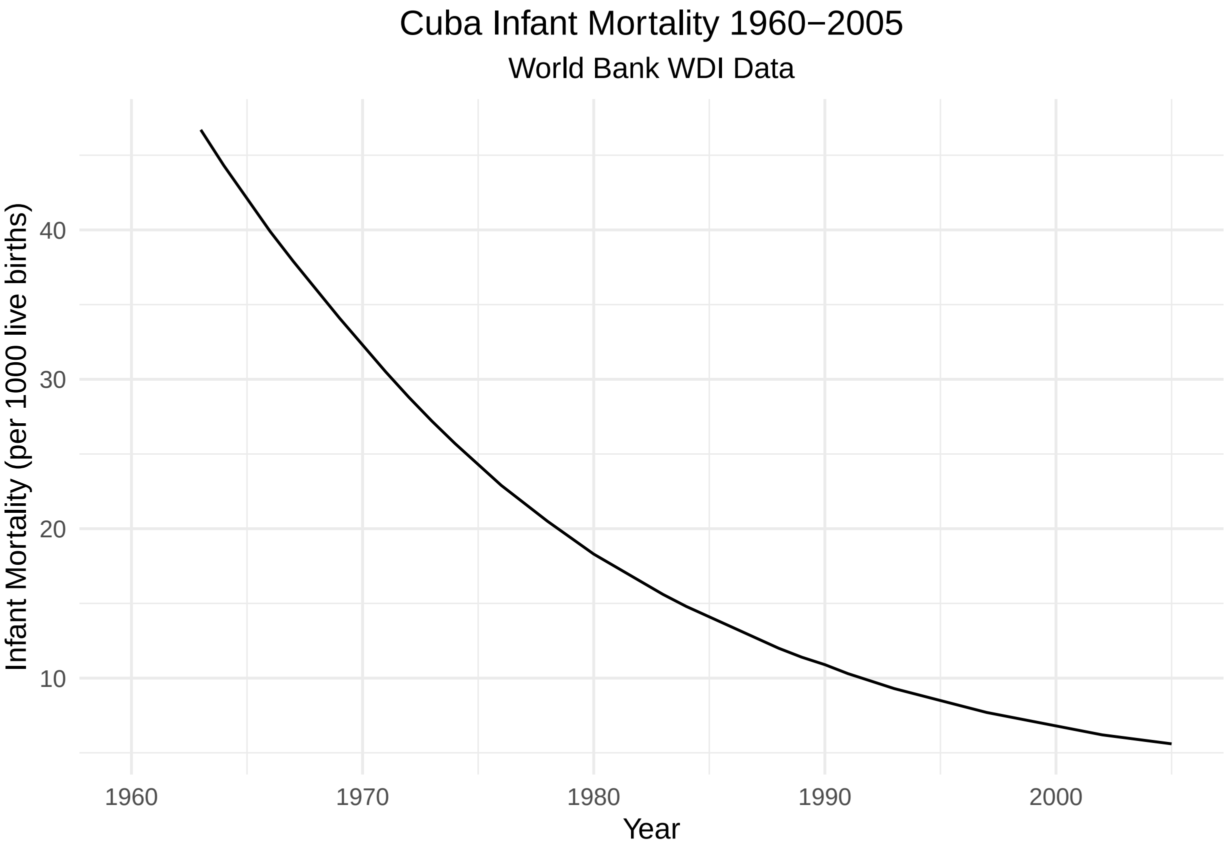
This groupwork assignment deals with multivariate plots. The dataset we will use is `wdi_development_data.csv`, which can be found in the Training Data folder. The dataset is a compilation of developmental indicators from the World Bank WDI data for 202 countries between the years 1960-2005.

For each of your graphs, please include a title and a subtitle and make sure to label your axes. Seek to make your graph as informative as possible: the goal is to have a plot that can (more or less) stand on its own!

```
# Each person's setwd() function will look a little different,
# depending on where they have saved the training data folder.
# It may look like:
# setwd("/Volumes/GoogleDrive/My Drive/Training Data August 2021")

library(tidyverse)
library(directlabels)
```

Exercise 1: For a country of your choosing, plot the infant mortality (given per 1000 life births) against time for the years 1960-2005 in a line graph. Briefly explain what trend you observe in the infant mortality over time.

```
# Note: The following example looks at Cuba. You can substitute it
  # in the code with your country of interest

df <- read.csv('wdi_development_data.csv') # load in the csv file

ggplot(subset(df, country %in% c("Cuba")), # subset so only include rows with country
      # == Cuba. All other data points excluded
       aes(x = year, y = inf_mort_WDI)) +
  geom_line() + # generating the line plot
  labs(title = "Cuba Infant Mortality 1960-2005",
       subtitle = "World Bank WDI Data",
       x = "Year",
       y = "Infant Mortality (per 1000 live births)") + # graph and axes labels
  theme_minimal() + # set background to white

  # Bonus adjustments to the plot
  theme(plot.title = element_text(hjust = .5)) + # centre the title
  theme(plot.subtitle = element_text(hjust = .5)) # centre the subtitle
```

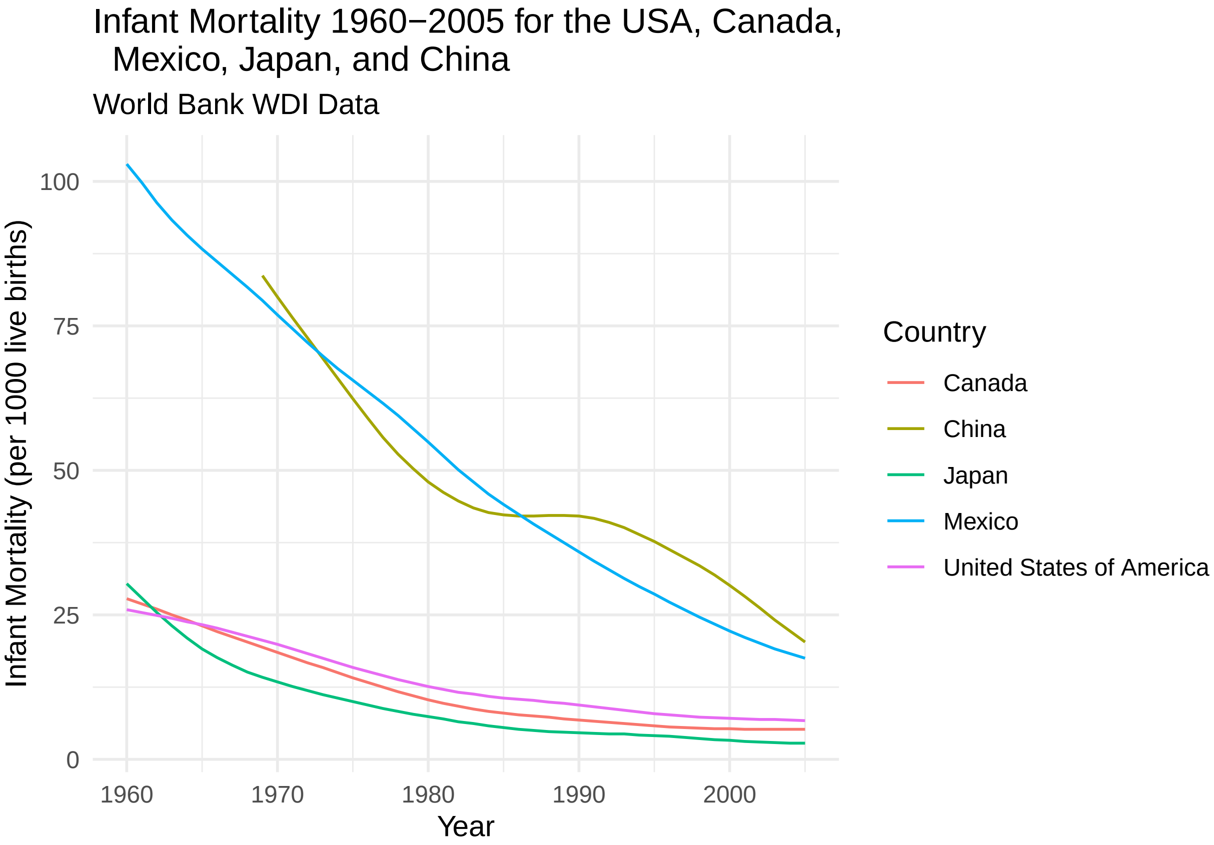


```
# Brief interpretation: There has been a steep decrease in infant mortality
  # since the 1960s, and appears to be gradually plateauing since the 2000s.
```

Exercise 2: For five countries of your choosing, plot the infant mortality against time for the years 1960-2005 in a line graph. Distinguish each country by color and make sure to include a legend. Comment on the differences across the 5 countries with time.

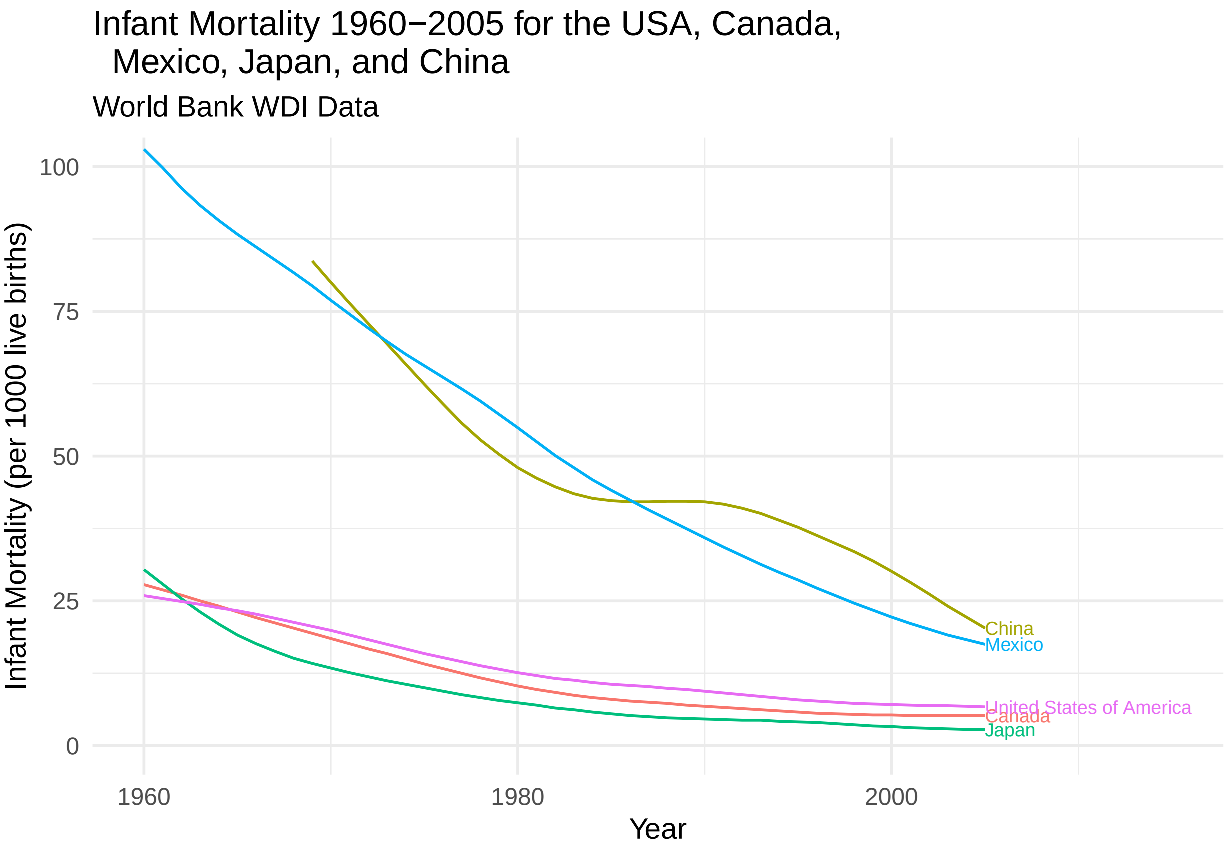
BONUS: Instead of a legend, add the names of the five countries as labels on the plot. This will require locating labels at specific coordinates that make it clear which line the label refers to. Note that this will only be possible if the countries selected have Y-axis values that are sufficiently distinct.

```
# As in the previous exercise, you can substitute the list
  # in the code with your five countries of interest.

library(directlabels) # for the country labels

ggplot(subset(df, country %in% c("United States of America", "Canada", "Mexico",
                                 "Japan", "China")),
       aes(x = year, y = inf_mort_WDI, group = country, label = country,
           color = factor(country))) +
  geom_line() +
  labs(title = "Infant Mortality 1960-2005 for the USA, Canada,
  Mexico, Japan, and China",
       # '\n' can be added to put the following text (after `\n`) in the next line
       subtitle = "World Bank WDI Data",
       x = "Year",
       y = "Infant Mortality (per 1000 live births)",
       color = "Country") + # this is to change the legend title
  theme(plot.title = element_text(hjust = .5)) +
  theme(plot.subtitle = element_text(hjust = .5)) +
  theme_minimal()
```



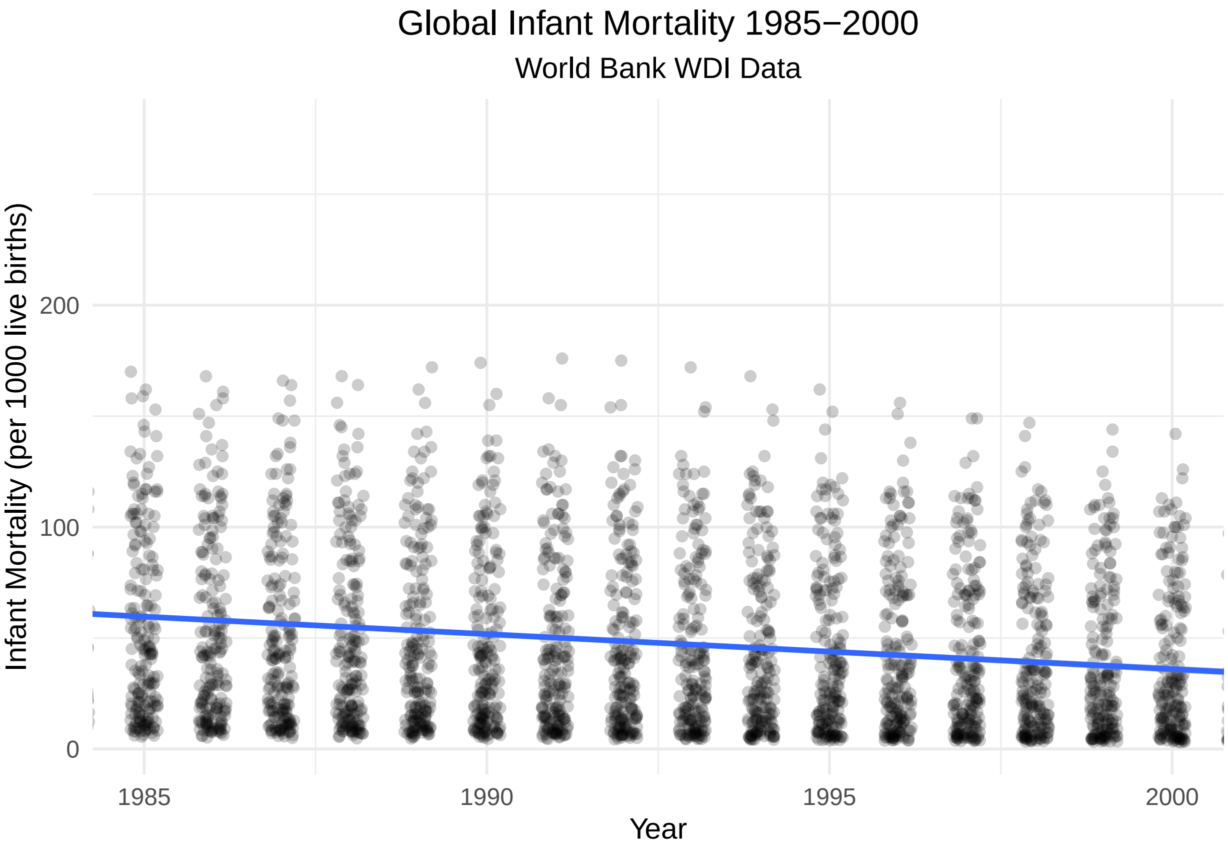
```
# Brief interpretation: There has been a steep decrease in infant mortality
  # since both Mexico and China, although China's decline has been more gradual.
  # Infant mortality have started at lower rates in the US, Canada, and Japan,
  # and so there decrease has been more gradual but have remained low for decades.
```

```
# Q2 BONUS answer

set.seed(1234)
ggplot(subset(df, country %in% c("United States of America", "Canada",
                                 "Mexico", "Japan", "China")),
       aes(x = year, y = inf_mort_WDI, group = country, label = country,
           color = country)) +
  labs(title = "Infant Mortality 1960-2005 for the USA, Canada,
  Mexico, Japan, and China",
       subtitle = "World Bank WDI Data",
       x = "Year",
       y = "Infant Mortality (per 1000 live births)") +
  theme(plot.title = element_text(hjust = .5)) +
  theme(plot.subtitle = element_text(hjust = .5)) +
  geom_line() +
  geom_dl(aes(label = country), method = list("last.points", cex = 0.55)) +
    # cex defines size of label as percent of label font size
  theme_minimal() +
  theme(legend.position = "none") +
  coord_cartesian(xlim = c(1960, 2015),
                  ylim = c(0, 100)) # adjusted x and y axes to allow room for labels
```



Exercise 3: In a scatterplot, plot infant mortality, for all countries, against time for the years 1985-2000. Include a jitter, choose an appropriate opacity, and include a trend line in your graph. What are the general trends that you observe in infant mortality? How does adjusting the opacity improve how informative your graph is?

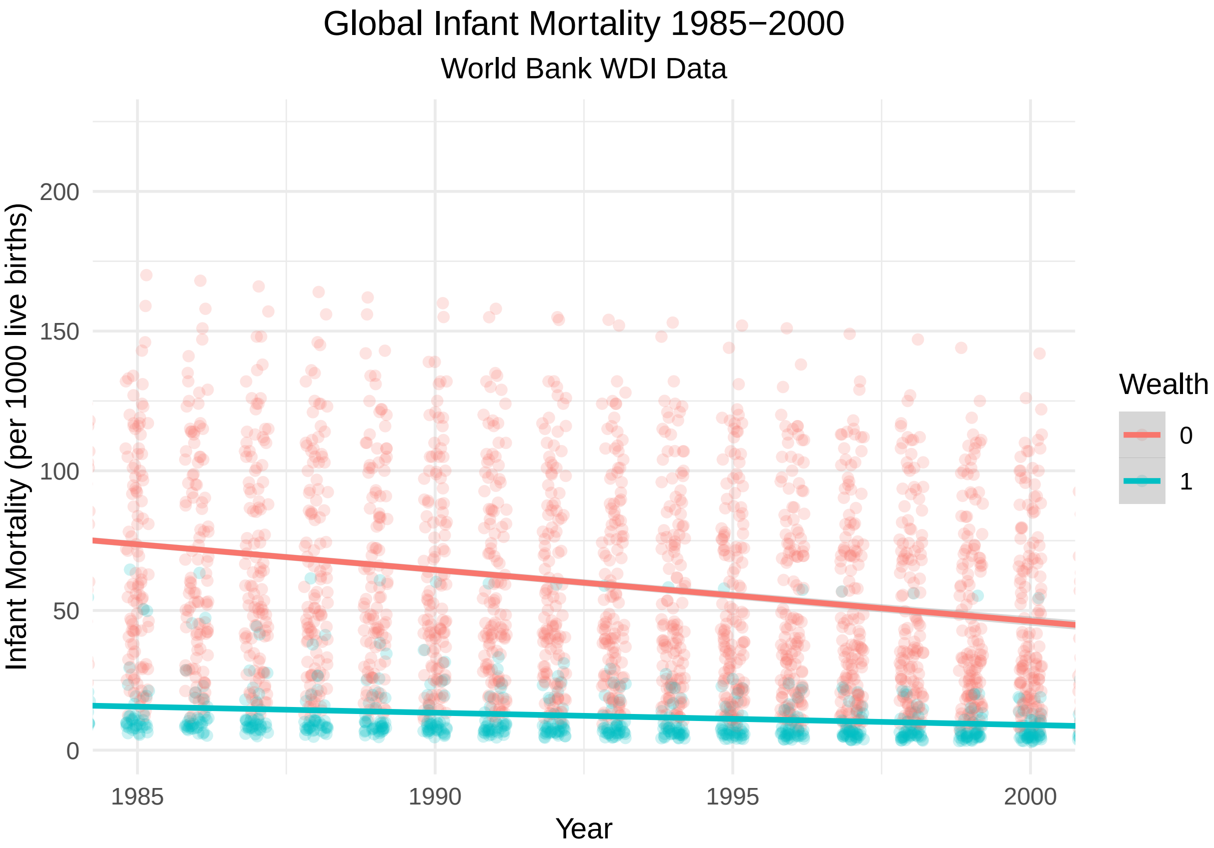
```
ggplot(df, aes(x = year, y = inf_mort_WDI)) +
  geom_point(alpha = 0.2, position = position_jitter(width = 0.2)) + # adjusting
    # opacity and jitter
  theme_minimal() +
  labs(title = "Global Infant Mortality 1985-2000",
       subtitle = "World Bank WDI Data",
       x = "Year",
       y = "Infant Mortality (per 1000 live births)") +
  theme(plot.title = element_text(hjust = .5)) +
  theme(plot.subtitle = element_text(hjust = .5)) +
  stat_smooth(method = "lm") + # including trend line
  coord_cartesian(xlim = c(1985, 2000))
```



```
# Brief interpretation: Globally, it appears that there is a steady decline
  # in infant mortality. Adjusting opacity allows us to better differentiate
  # individual points, and see where countries are clustered in terms of rates of
  # infant mortality. Many countries are clustered at the bottom, or relatively
  # lower rates of infant mortality. There are fewer observations
  # at the highest levels of infant mortality on the global scale.
```

Exercise 4: Lastly, create a dummy variable called wealthy. Classify countries with a gdppc > mean gdppc as wealthy (dummy = 1) and those with a gdppc < mean gdppc as not wealthy (dummy = 0). Distinguish data points in your graph from exercise 3 using this dummy variable using both shape and color, so that there is a clear distinction between the data points from the two groups of countries respectively. Comment on trends across time.

```
df$wealthy <- ifelse(df$gdppc_WDI >= mean(df$gdppc_WDI, na.rm = T), 1, 0)
  # creating dummy
table(df$wealthy) # generates number of values in each dummy variable category

ggplot(subset(df, !is.na(wealthy)), # choosing only non-null values here
       aes(x = year, y = inf_mort_WDI,
           color = factor(wealthy))) +
  geom_point(alpha = 0.2, position = position_jitter(width = 0.2)) +
  theme_minimal() + # white background
  labs(title = "Global Infant Mortality 1985-2000",
       subtitle = "World Bank WDI Data",
       x = "Year",
       y = "Infant Mortality (per 1000 live births)",
      color = "Wealth") +
  theme(plot.title = element_text(hjust = .5)) +
  theme(plot.subtitle = element_text(hjust = .5)) +
  stat_smooth(method = "lm") +
  coord_cartesian(xlim = c(1985, 2000)) # adjust x-axis range of values
```



```
# Brief interpretation: Similar to the previous graph, globally, there is a
  # steady decline in infant mortality. Differentiating between wealthy and
  # non-wealthy countries, we can see big differences in trends in infant mortality.
  # Wealthy countries are clustered on the lower end of infant mortality rates,
  # and have remained low since the 1960s. While non-wealthy countries are more
  # spread out, and relatively higher in terms of infant mortality rates,
  # we can see a clearer decrease over time.
```