SPEC REU R Resources: Multivariate Plots – Group Work

Radhika Ananth, Gaea Morales, Claudia Salas Gimenez and Ben Graham

Summer 2024

Welcome to the groupwork assignment on multivariate plots! In this document, you'll apply what you've learned in the previous four walkthroughs on multivariate plots by visualizing data trends across countries using key developmental indicators.

For each graph, be sure to include a title, subtitle, and appropriately labeled axes to clearly convey the information on its own. Additionally, challenge yourself to enhance the plot with extra aesthetic elements for high-quality, polished visuals.

Initial Setup

Before we dive into the exercises, set up your environment and load the necessary libraries and dataset. For this assignment, we'll use a compilation of developmental indicators from the World Bank WDI data for 202 countries from 1960 to 2005, saved as `wdi_development_data.csv`.

```
# Set working directory
#setwd("YourFolderPath")

# Load required libraries
library(tidyverse)
library(directlabels)
library(readr)

# Load the dataset
df <- read_csv("wdi_development_data.csv")
```

For more details on the indicators used, check the World Development Indicators data catalog.

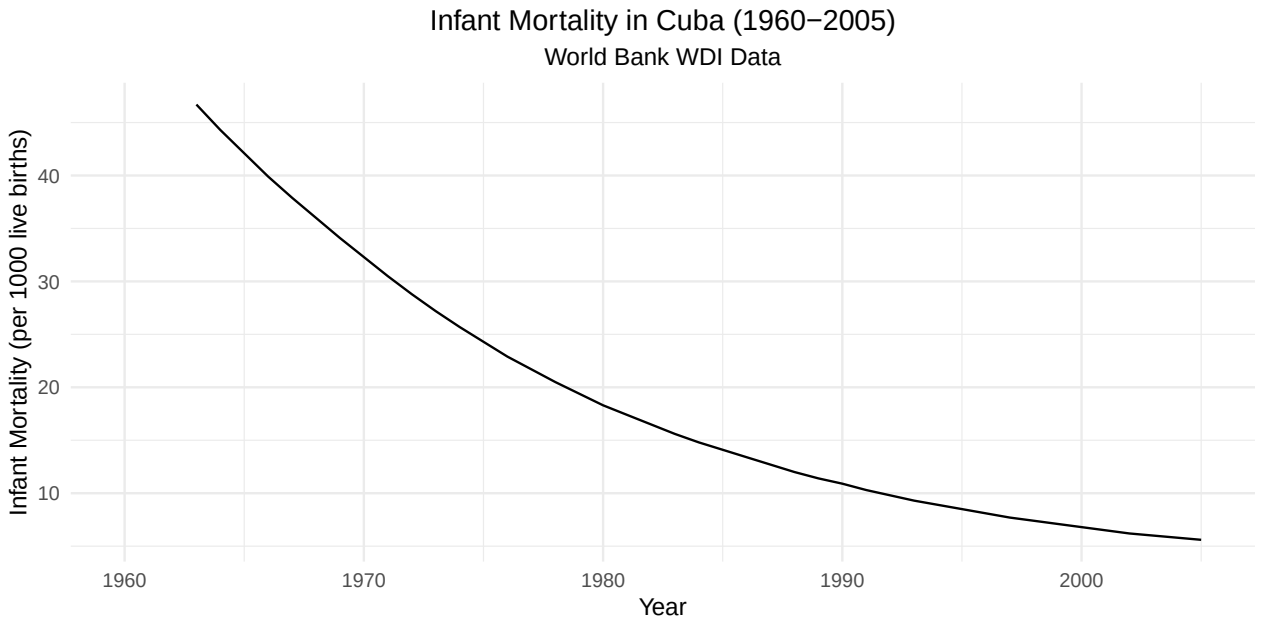
Exercise 1: Plotting Infant Mortality for a Selected Country

Select a country and create a line plot showing infant mortality rates (per 1,000 live births) over time, from 1960 to 2005. Then, briefly interpret the trends for the country you selected.

```
# For the answer key, we'll use Cuba to plot infant mortality, so keep in mind that
# the answer key might look slightly different for other countries.

# Create the plot for Cuba
ggplot(subset(df, country %in% c("Cuba")),
       # Subset data to include only rows where country == Cuba
       aes(x = year, y = inf_mort_WDI)) +
  geom_line() + # Generate the line plot
  labs(title = "Infant Mortality in Cuba (1960-2005)",
       subtitle = "World Bank WDI Data",
       x = "Year",
```

```
      y = "Infant Mortality (per 1000 live births)") +
  # Add graph title, subtitle, and axis labels
  theme_minimal() + # Set background to white
  # Bonus adjustments to the plot
  theme(plot.title = element_text(hjust = .5)) +
  theme(plot.subtitle = element_text(hjust = .5))
```




```
  # Center-align the title and subtitle

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

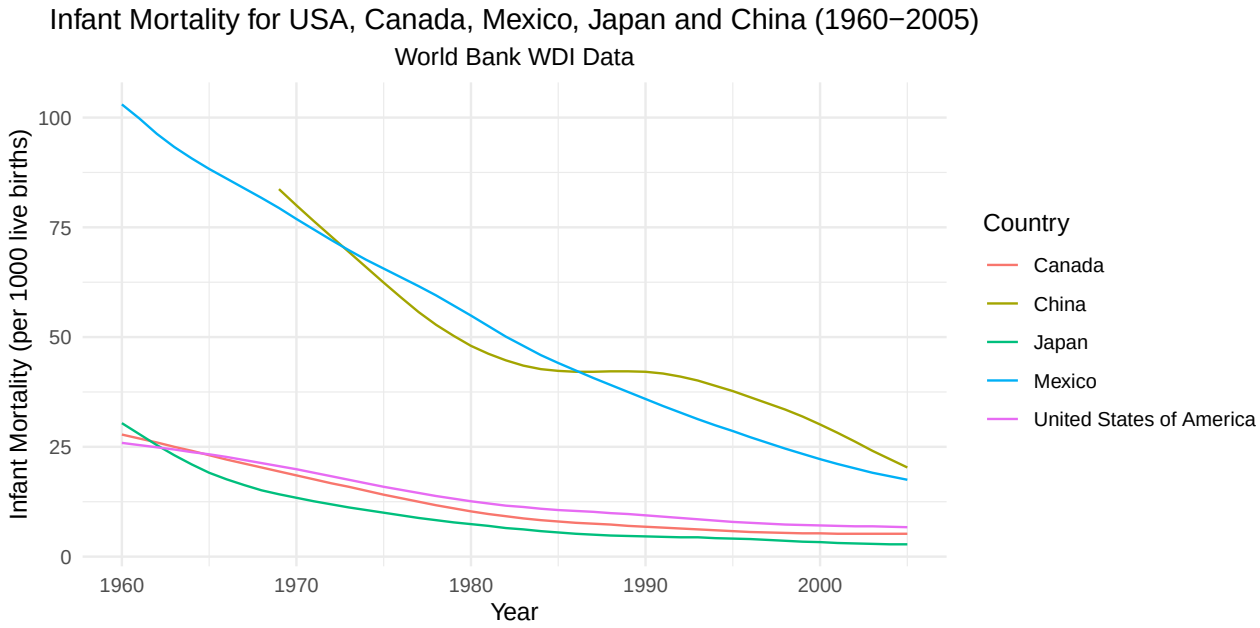
Exercise 2: Comparing Infant Mortality Across Five Countries

Choose five countries and compare their infant mortality rates over time (1960-2005). Differentiate each country by color, and include a legend. Comment on the observed trends.

```
# For the answer key, we'll use the United States, Canada, Mexico, Japan and China
# to plot infant mortality rates, so keep in mind that the answer key might look
# slightly different for other countries.

# Create the plot for selected countries
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 for USA, Canada, Mexico, Japan and China (1960-2005)",
       subtitle = "World Bank WDI Data",
       x = "Year",
       y = "Infant Mortality (per 1000 live births)",
       color = "Country") +
  # Add labels and title to the plot
  theme_minimal() +
```

```
  # Apply a minimal theme for a clean look
  theme(plot.title = element_text(hjust = .5),
        plot.subtitle = element_text(hjust = .5))
```




```
  # Center-align the title and subtitle

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

Bonus Question (Exercise 2)

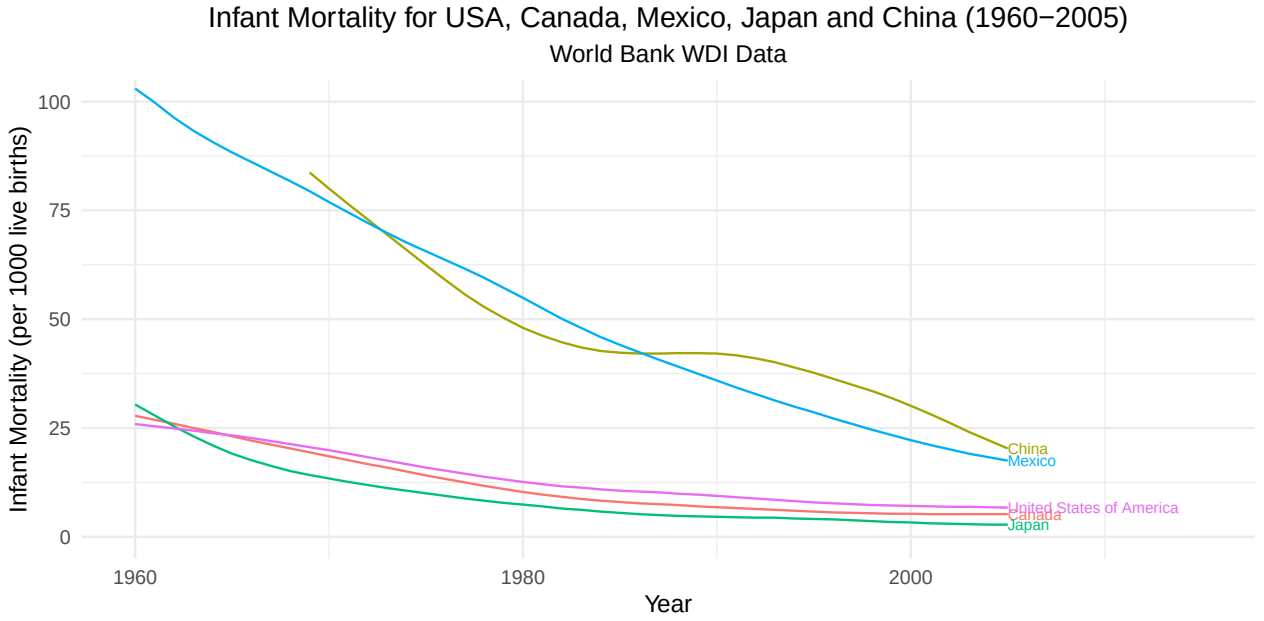
Add a label for each country directly on the plot instead of using a legend, if the trends allow clear differentiation.

Note: This will only be feasible if the selected countries have sufficiently distinct y-axis values.

```
# Set seed for reproducibility
set.seed(1234)

# Create the plot with direct 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 = country)) +
  geom_line() +
  labs(title = "Infant Mortality for USA, Canada, Mexico, Japan and China (1960-2005)",
       subtitle = "World Bank WDI Data",
       x = "Year",
       y = "Infant Mortality (per 1000 live births)") +
  theme_minimal() +
  geom_dl(aes(label = country), method = list("last.points", cex = 0.55)) +
  # Add direct labels to the plot
```

```
  theme(legend.position = "none") +
  # Remove the legend
  theme(plot.title = element_text(hjust = .5)) +
  theme(plot.subtitle = element_text(hjust = .5)) +
  coord_cartesian(xlim = c(1960, 2015),
                  ylim = c(0, 100))
```



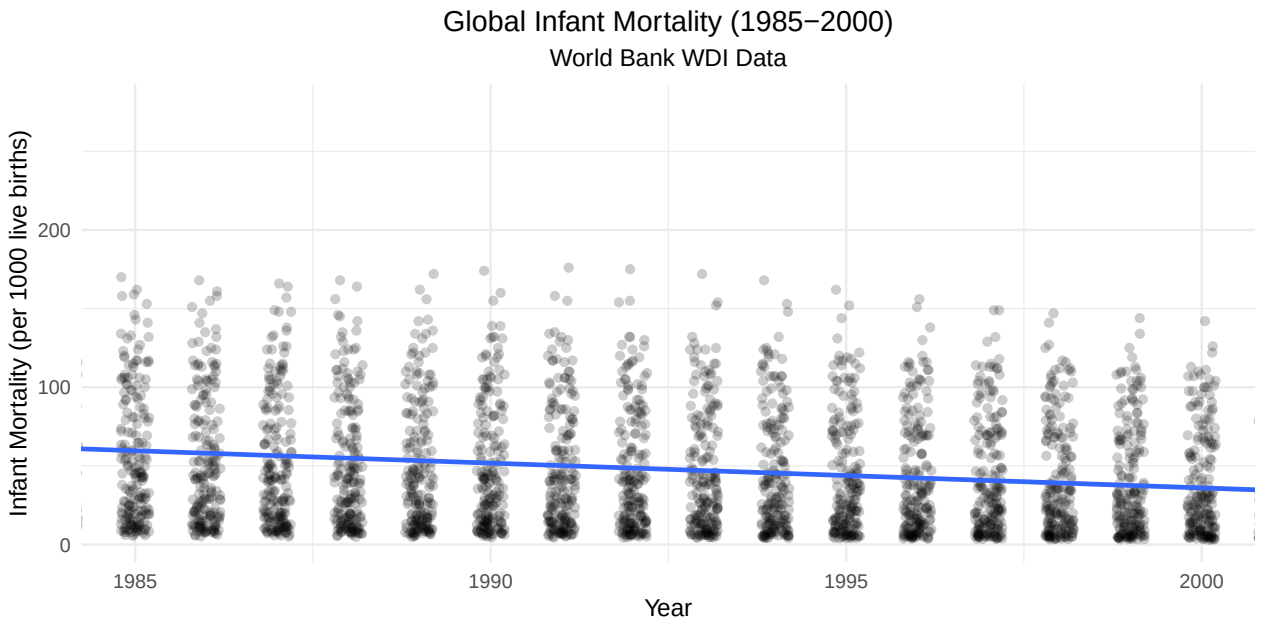

```
  # Adjust the plot's x and y axis limits to leave space for the country labels
```

Exercise 3: Global Trends in Infant Mortality

Create a scatterplot to display global infant mortality rates over time for the years 1985 to 2000. Adjust the opacity and add jitter to the points for clarity, and be sure to include a trend line to highlight overall trends. Feel free to incorporate additional aesthetic elements to make the visualization more engaging and informative.

What are the general trends that you observe in infant mortality? How does adjusting the opacity improve how informative your graph is?

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



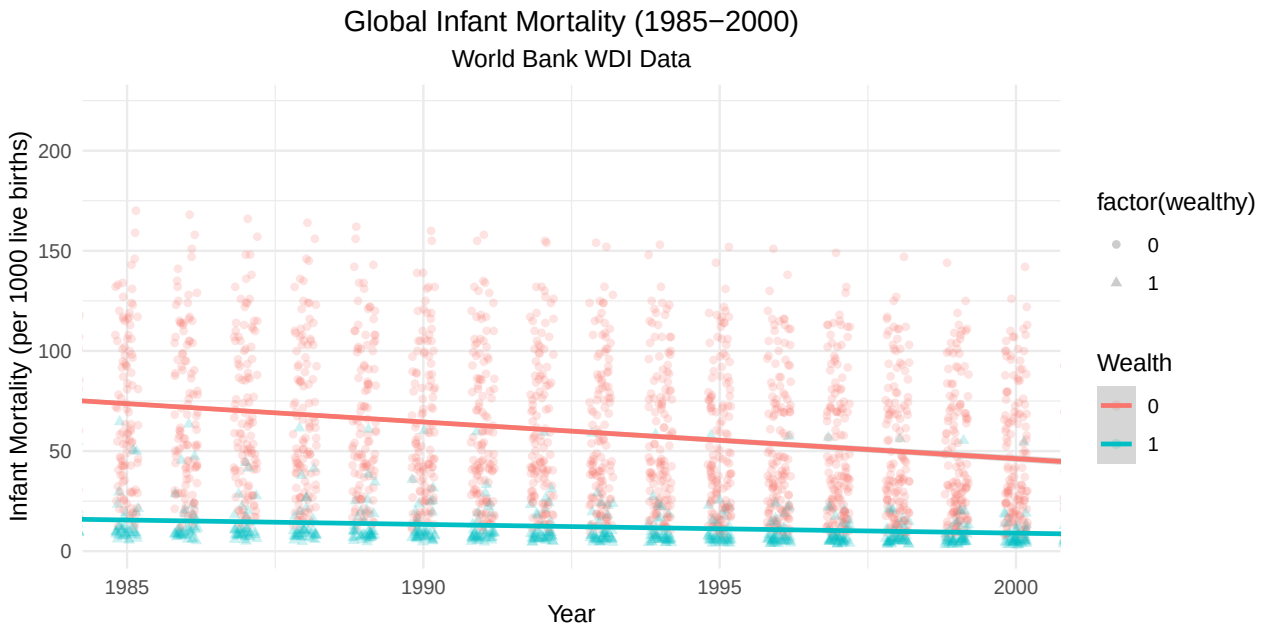

```
# Brief interpretation: Globally, there is a steady decline in infant mortality.
#Adjusting opacity allows us to better differentiate individual points and see where
# countries are clustered in terms of rates of infant mortality. Many countries are
# clustered at the bottom, with relatively lower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

Exercise 4: Impact of Wealth on Infant Mortality

For this last exercise, create a dummy variable called `wealthy` based on GDP per capita (`gdppc`), classifying countries as high-income if their `gdppc` is above the mean, assigning them a value of 1, while countries below the mean receive a 0. Use this `wealthy` variable to distinguish data points data points in the scatterplot from Exercise 3 by shape and color in a scatter plot, then interpret the observed trends.

```
# Create the dummy variable 'wealthy'
df$wealthy <- ifelse(df$gdppc_WDI >= mean(df$gdppc_WDI, na.rm = T), 1, 0)

# Create the scatter plot with the dummy variable
ggplot(subset(df, !is.na(wealthy)),
       # Choose non-null values only
       aes(x = year, y = inf_mort_WDI, color = factor(wealthy), shape = 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))
```




```
# 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.
```

Bonus Question

For an extra challenge, try creating a new visualization by selecting variables from the `wdi_development_data.csv` file or data from one of R's preloaded datasets (use `data()` to list them, and `?dataset_name` to see details about each). Choose a plot type that best represents your variables and aligns with the purpose of the figure. Additionally, include a brief description explaining why you chose this type of plot—what insights are you aiming to reveal, and how does the plot effectively highlight that information?