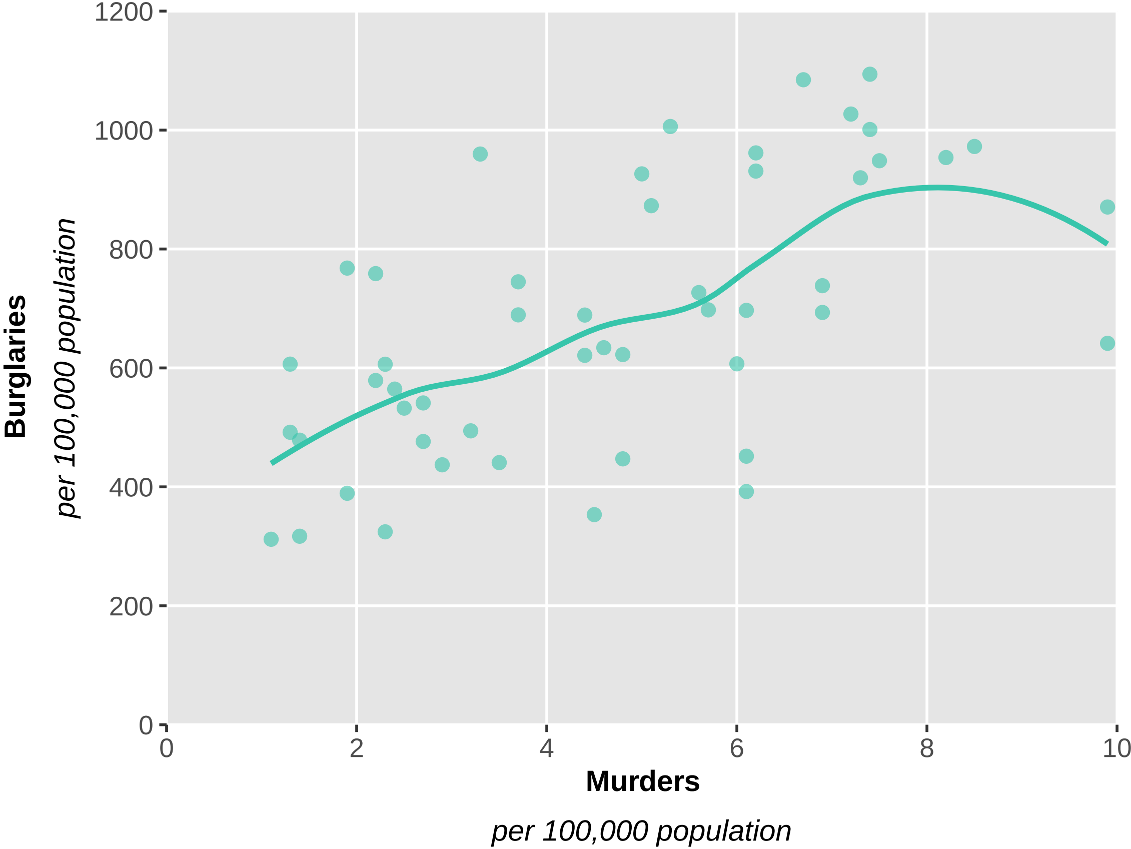
MURDERS VERSUS BURGLARIES IN THE UNITED STATES

States with higher murder rates tend to have higher burglary rates



Assignment 1 Task 1

```
library (dplyr)
library(ggplot2)
library(readr)
library(formatR)

crime_by_state_2005 <- read.csv("/Users/talakammar/Library/Mobile Documents/com~apple~CloudDocs/Spring 2025/Data
Visualization/Assignment 1/crimeRatesByState2005.csv", header = TRUE, sep = ",")

crime_filtered <- filter(crime_by_state_2005, murder <= 10)

ggplot(data = crime_filtered, mapping = aes(x = murder, y = burglary)) +
  geom_point(color = "#37c5ab", size = 2, alpha = 0.6) +
  geom_smooth(method = "loess", color = "#37c5ab", se = FALSE, size = 1) +
  labs(
    title = "MURDERS VERSUS BURGLARIES IN THE UNITED STATES",
    subtitle = expression(italic("States with higher murder rates tend to have higher burglary rates")),
    x = expression(atop(bold("Murders"), italic("per 100,000 population"))),
    y = expression(atop(bold("Burglaries"), italic("per 100,000 population"))))+
  scale_x_continuous(
    limits = c(0, 10),
    breaks = seq(0, 10, by = 2),
    expand = c(0, 0)) +
  scale_y_continuous(
    limits = c(0, 1200),
    breaks = seq(0, 1200, by = 200),
    expand = c(0, 0)) +

  theme(
    panel.background = element_rect(fill = "grey90", color = NA),
    panel.grid.minor = element_blank(),
    plot.title = element_text(hjust = 0.5, face = "bold", size = 13),
    plot.subtitle = element_text(hjust = 0.5, size = 11,  vjust = 2 ),

    axis.title.x = element_text(size = 11),
    axis.title.y = element_text(angle = 90, size = 11),

    axis.text.x = element_text(size = 10),
    axis.text.y = element_text(size = 10))
```

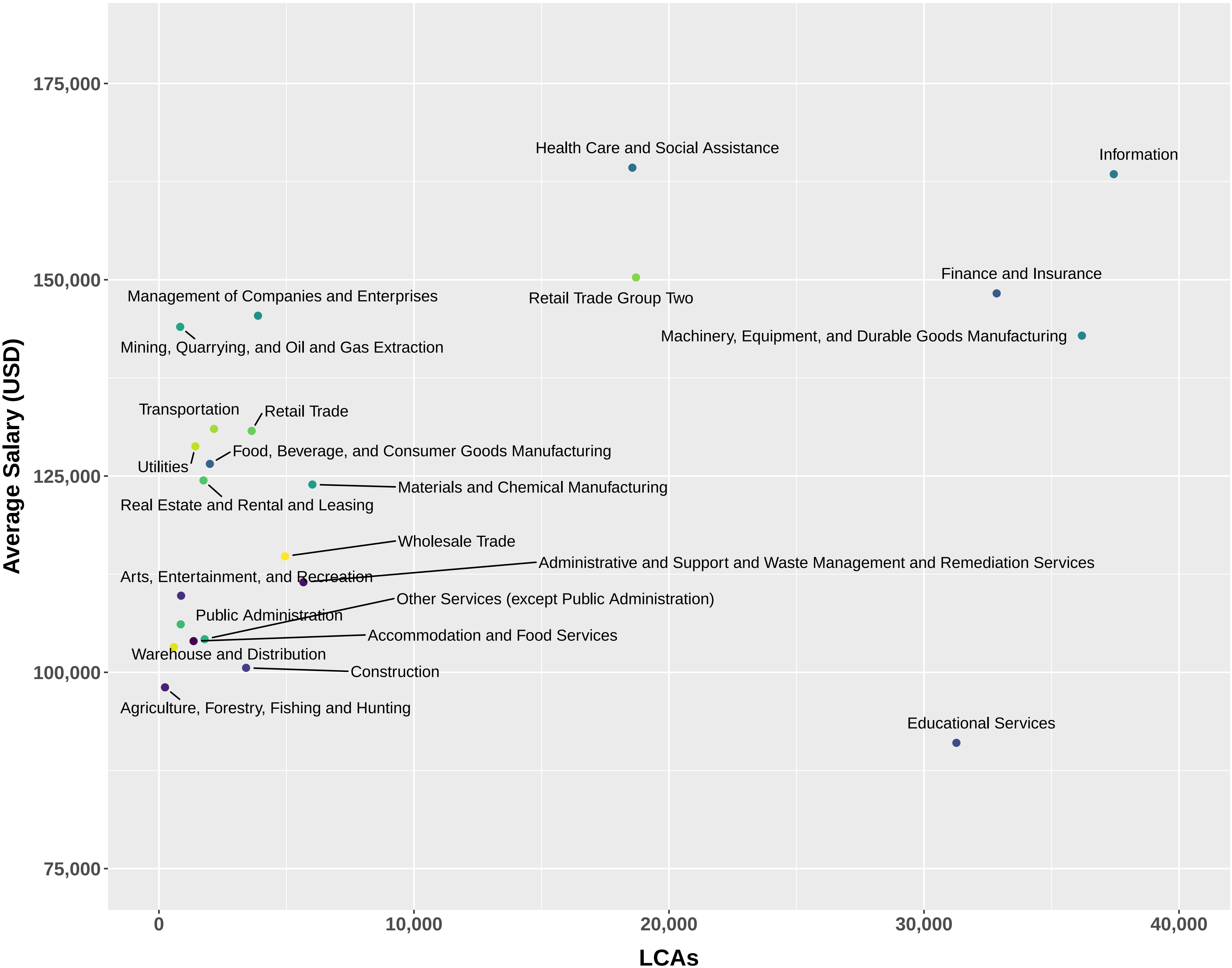
Assignment 1 Task 2

Statement of Purpose

The scatterplot below explores how the volume of Labor Condition Applications (LCAs) compares to the average salaries in different industries. A LCA is an essential step in the H1B visa process in the United States. In order for employers to hire someone on an H-1B visa, they must first file an LCA with the Department of Labor to confirm that the job's pay either meets or exceeds the prevailing wage. By plotting LCAs on the x-axis and average salaries on the y-axis, I'm looking to see if industries with more LCAs also tend to offer higher pay. Beyond illustrating this relationship, the visualization helps identify which sectors attract the most H1B applicants and whether that high demand correlates with higher compensation.

Labor Condition Applications (LCAs) VS. Salaries by Industry

Industries with High LCAs have predominantly High Salaries



Assignment 1 Task 2

```
library(dplyr)
library(ggplot2)
library(readr)
library(ggrepel)
library(readxl)

H1B_by_Industry <- read_excel("~/Library/Mobile Documents/com~apple~CloudDocs/Spring 2025/Data Visualization/Assignment 1/Top H1-B Visas by Industry.xlsx")

colnames(H1B_by_Industry) <- c("Rank","NAICS_Industry", "Number_of_LCA", "Average_Salary")

H1B_filtered <- select(H1B_by_Industry, -Rank)
H1B_filtered_2 <- filter(H1B_filtered, Number_of_LCA < 200000)
```

```
ggplot(data = H1B_filtered_2,
       mapping= aes(x = Number_of_LCA, y = Average_Salary)) +
  geom_point(aes(color = NAICS_Industry), size = 2) +
  geom_text_repel(aes(label = NAICS_Industry), size = 3.7,
                  box.padding = unit(0.6, "lines"),
                  point.padding = unit(0.6, "lines")) +
  scale_color_viridis_d(guide = "none") +
  labs(
    title = "Labor Condition Applications (LCAs) VS. Salaries by Industry",
    subtitle = expression(italic("Industries with High LCAs have predominantly High Salaries")),
    x = expression(bold("LCAs")),
    y = expression(bold("Average Salary (USD)"))
  ) +
  scale_x_continuous(
    labels = scales::comma,
    limits = c(0, 40000)) +
  scale_y_continuous(
    labels = scales::comma,
    limits = c(75000, 180000))+
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold", size = 30),
    plot.subtitle = element_text(hjust = 0.5, size = 16,  vjust = 2),
    axis.text.x = element_text(size = 13, face = "bold"),
    axis.text.y = element_text(size = 13, face = "bold"),
    axis.title.x = element_text(size = 16, vjust = -1, face = "bold"),
    axis.title.y = element_text(size = 16, face = "bold"))
```

Sources:

- I used the following website to findout how to make a subtitle for my axes titles: https://stackoverflow.com/questions/54093292/how-to-add-subtext-to-axes-in-ggplot2-r

- I used the following website to learn how to change positions of axes titles: https://stackoverflow.com/questions/75489579/change-axis-title-position-in-ggplot2

- Pulled the color hex code for my aesthetics from the following website: https://htmlcolorcodes.com/

- used the following website to edit axes limits and tick marks: https://stackoverflow.com/questions/44170871/how-does-ggplot-scale-continuous-expand-argument-work

- used the following source to find out how to space out dot labels/ change padding https://www.rdocumentation.org/packages/ggrepel/versions/0.9.6/topics/geom_label_repel
- used the following source to check methods of manipulating minor ticks: https://www.tidyverse.org/blog/2024/02/ggplot2-3-5-0-axes/
- to avoid overlapping texts in labels: https://ggrepel.slowkow.com/articles/examples.html

AI prompts and code used:

- ***"how do i make bigger grid squares in the following code so that the background squares are larger:***

```
library (dplyr)
library(ggplot2)
library(readr)

Crime_Rates_State_2005 <- read.csv("/Users/talakammar/Library/Mobile Documents/com~apple~CloudDocs/Spring 2025/Data Visualization/Assignment 1/crimeRatesByState2005.csv", header = TRUE, sep = ",")

crime_filtered <- filter(Crime_Rates_State_2005, murder <= 10)

ggplot(data = crime_filtered, mapping = aes(x = murder, y = burglary)) +
  geom_point(color = "#37c5ab", size = 2, alpha = 0.6) +
  geom_smooth(method = "loess", color = "#37c5ab", se = FALSE, size = 1) +
  labs(
    title = "MURDERS VERSUS BURGLARIES IN THE UNITED STATES",
    subtitle = expression(italic("States with higher murder rates tend to have higher burglary rates")),
    x = expression(atop(bold("Murders"), italic("per 100,000 population"))),
    y = expression(atop(bold("Burglaries"), italic("per 100,000 population"))))”
```

Lines used:

```
scale_x_continuous(
limits = c(0, 10),
breaks = seq(0, 10, by = 2),
expand = c(0, 0)) +
scale_y_continuous(
limits = c(0, 1200),
breaks = seq(0, 1200, by = 200),
expand = c(0, 0))
```

- **“*in the scatter plot i made below, how do i make the dots be labeled as my industry names which is my third variable?***

```
library (dplyr)
ibrary(ggplot2)
library(readr)


H1B_filtered <- select(H1B_by_Industry, -Rank)
H1B_filtered_2 <- filter(H1B_filtered, Number of LCA *<200000)

H1B_filtered_2

colnames(H1B_filtered_2) <- c("NAICS_Industry", "Number_of_LCA",
Average_Salary")


ggplot(data = H1B_filtered_2, mapping = aes(x = Number_of_LCA, y =
Average_Salary)) +
  geom_point(size = 2, alpha = 0.6
  labs(
    title = "Number of H1-B Visas Issued Vs. Average Salary",
    subtitle = expression(italic("Industries with higher LCA numbers tend to have
higher salaries")),
    x = "Number of LCAs",
    y = "Average Salary",
    color = "NAICS Industry")”
```

Lines used:

```
- geom_text_repel(aes(label = NAICS_Industry), size = 3.7)+
  scale_color_viridis_d(guide = "none")
```

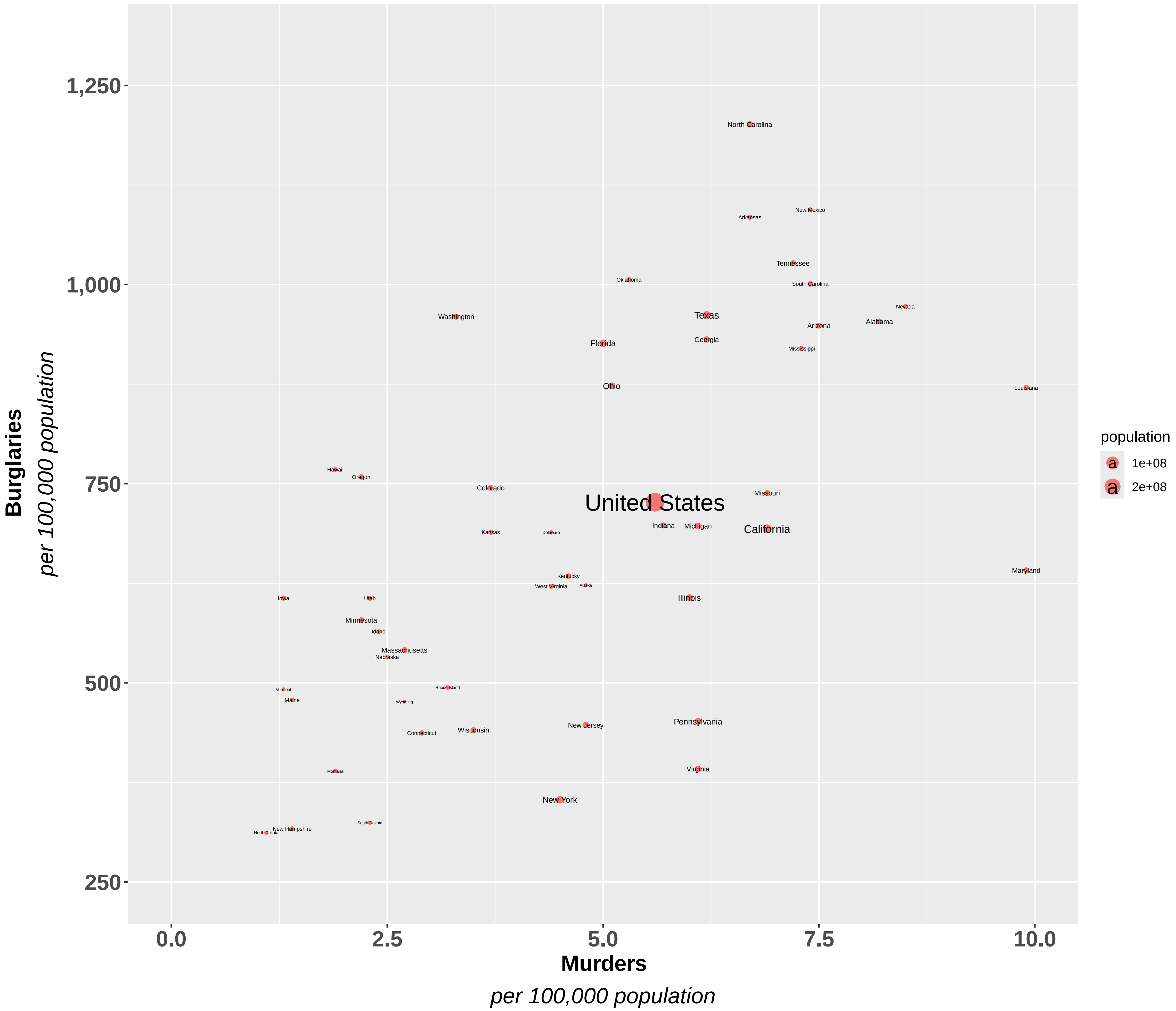
- ***“how do I add multiple lines as layers to my line plot in “geom_line” as per my code:***

```
library(readxl)
library(ggplot2)

migrant_stock <- read_excel("/Users/talakammar/Library/Mobile
Documents/com~apple~CloudDocs/Spring 2025/Data Visualization/Data set on
migration/undesa_pd_2024_ims_stock_by_sex_and_origin.xlsx",
                    sheet = "Table 1",
                    range = "A11:M299")

migrant_stock <- migrant_stock[-
c(1:27,48,58,66,72,90,91,97,116,106,128,147,148,159,173,190,200,230,239,
                         254,260,288), -c(1,3,4,5)]
migrant_stock <- migrant_stock[-c(158),]

View(migrant_stock)
colnames(migrant_stock) <- c("Country", "1990", "1995", "2000", "2005", "2010",
"2015",
                    "2020", "2024")

countryNames <- migrant_stock$Country
migrant_stock_t <- as.data.frame(t(migrant_stock[,-1]))
colnames(migrant_stock_t) <- countryNames
migrant_stock_t$Year <- as.numeric(rownames(migrant_stock_t))

total_migrants_per_country <- colSums(migrant_stock_t, na.rm = TRUE)
sorted_migrants<- order(total_migrants_per_country, decreasing = TRUE)
names_top5Countries <-
names(total_migrants_per_country)[sorted_migrants][1:5]
top5countries <- migrant_stock_t[, c(names_top5Countries, "Year")]


ggplot(data = top5countries, aes(x = Year)) +
geom_line()
labs(title = "Top 5 Countries by Migrant Stock (1990–2024)",
  subtitle = expression(italic("All 5 countries except Russia have a higher number
of
                          migrants compared to their 1990 levels")),
  x = "Year",
  y = "Number of Migrants (in millions)") “
```

- **Line used:**

```
geom_line(aes(y = !!sym(names_top5Countries[1]),color =
names_top5Countries[1]),size = 1)
```

MURDERS VERSUS BURGLARIES IN THE UNITED STATES



Source: U.S. Census Bureau | Nathan You

Assignment 2 Task 1

```r
library(readr)
library(ggplot2)
library(readr)
library(ggrepel)


crime_by_state_2005 <- read.csv("/Users/talakammar/Library/Mobile Documents/com~apple~CloudDocs/Spring 2025/Data Visualization/Assignment 1/crimeRatesByState2005.csv", header = TRUE, sep = ",")
```

```r
ggplot(data = crime_by_state_2005,
       mapping = aes(x = murder,
                     y = burglary,
                     size = population)) +
  geom_point(shape=16, alpha = 0.5, color = "red") +
  geom_text(aes(label = state, show.legend = FALSE)) +
labs(
  title = "MURDERS VERSUS BURGLARIES IN THE UNITED STATES",
  x = expression(atop(bold("Murders"), italic("per 100,000 population"))),
  y = expression(atop(bold("Burglaries"), italic("per 100,000 population"))),
  caption = "Source: U.S. Census Bureau | Nathan You") +

  scale_x_continuous(
    labels = scales::comma,
    limits = c(0, 10)) +
  scale_y_continuous(
    labels = scales::comma,
    limits = c(250, 1300 ))+

  theme(
    plot.title = element_text(vjust = 2, hjust = 0.5, face = "bold",
                              size = 25),
    plot.subtitle = element_text(hjust = 0.5, size = 11,  vjust = 2 ),
    axis.text.x = element_text(size = 16, face = "bold"),
    axis.text.y = element_text(size = 16, face = "bold"),
    axis.title.x = element_text(size = 16, face = "bold"),
    axis.title.y = element_text(size = 16, face = "bold"))
```

Sources:

- **Used to insert citation into plot:**

https://stackoverflow.com/questions/10197738/add-a-footnote-citation-outside-of-plot-area-in-r

https://stackoverflow.com/questions/6778908/transpose-a-data-frame

Assignment 3 Task 1

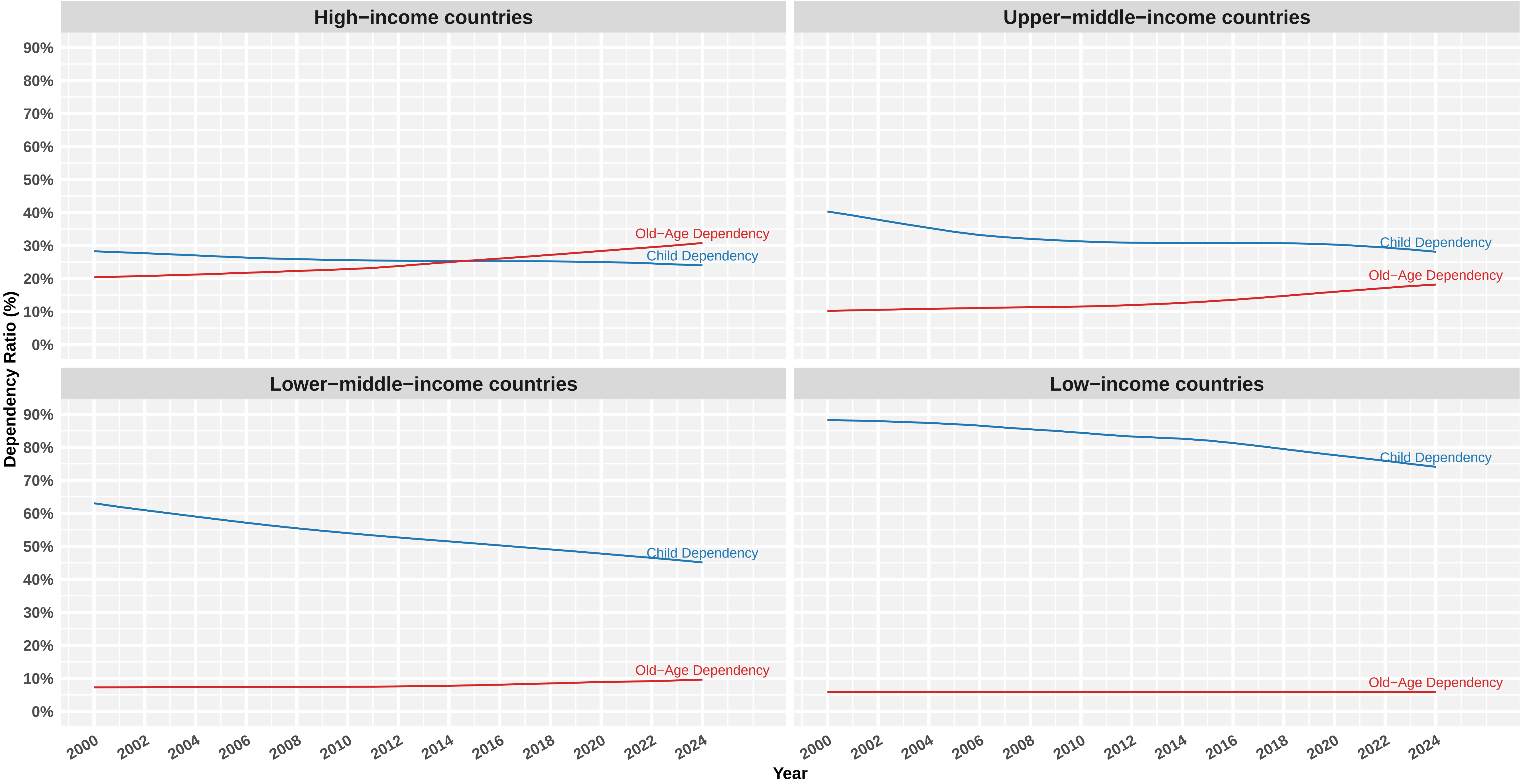
Statement of purpose

In this chart, I visualize how child and old-age dependency ratios have changed from 2000 to 2024 across four global income groups. By plotting both types of dependency for each region, I aim to show how the demographic burden of supporting younger and older populations is shifting. Child dependency has fallen, especially in low-income countries, while old-age dependency has steadily risen in high-income nations. This comparison helps reveal how countries at different development levels are preparing for aging populations or transitioning away from high youth burdens.

Dependency Trends by Income Group (2000−2024)

Across all income groups, child dependency has declined since 2000−most notably in low−income countries. Meanwhile, old−age dependency has risen steadily, especially in high−income and upper−middle−income countries. By 2024, the burden of supporting older adults is nearly as high as that of children in wealthier nations. These shifting patterns underscore the different demographic pressures countries face depending on their level of development.

Source: UN Population Division, World Population Prospects (2024).



Assignment 3 Task 1

2025-04-23

R Markdown

```r
library(dplyr)
library(ggplot2)
library(tidyr)
library(scales)

dependancy_data <- read.csv("/Users/talakammar/Library/Mobile Documents/com~apple~CloudDocs/Spring 2025/Data Visualization/unpopulationdata_DependancyRatio.csv")

# Filter and prep
df_filtered <- dependancy_data %>%
  filter(
    IndicatorShortName %in% c("Child dependency ratio", "Old-age dependency ratio"),
    Location %in% c("Low-income countries", "Lower-middle-income countries",
                    "Upper-middle-income countries", "High-income countries"),
    Time >= 2000, Time <= 2024,
    EstimateType == "Model-based Estimates"
  ) %>%
  select(Time, Location, IndicatorShortName, Value) %>%
  rename(Ratio = IndicatorShortName) %>%
  mutate(
    Ratio = recode(Ratio,
                   "Child dependency ratio" = "Child Dependency",
                   "Old-age dependency ratio" = "Old-Age Dependency"),
    Location = factor(Location, levels = c(
      "High-income countries",
      "Upper-middle-income countries",
      "Lower-middle-income countries",
      "Low-income countries"))
  )

df_labels <- df_filtered %>%
  filter(Time == 2024)

ggplot(data = df_filtered, aes(x = Time, y = Value, color = Ratio)) +
  geom_line(size = 0.6) +
  geom_text(data = df_labels, aes(label = Ratio, vjust = -0.5, hjust = 0.5),
            size = 3, show.legend = FALSE) +
  facet_wrap(~Location, ncol = 2) +
  scale_y_continuous(
    limits = c(0, 90),
    breaks = seq(0, 90, by = 10),
    labels = label_percent(scale = 1)
  ) +
  scale_x_continuous(
    limits = c(2000, 2026),
    breaks = seq(2000, 2024, by = 2)
  ) +
  scale_color_manual(
    values = c("Child Dependency" = "#1f77b4", "Old-Age Dependency" = "#d62728")
  ) +
  labs(
    title = "Dependency Trends by Income Group (2000-2024)",
    subtitle = "
Across all income groups, child dependency has declined since 2000—most notably in low-income countries. Meanwhile, old-age dependency has risen steadily, especially in high-income
and upper-middle-income countries. By 2024, the burden of supporting older adults is nearly as high as that of children in wealthier nations. These shifting patterns underscore the
different demographic pressures countries face depending on their level of development.

Source: UN Population Division, World Population Prospects (2024).",
    x = "Year",
    y = "Dependency Ratio (%)"
  ) +
  theme_minimal(base_size = 11) +
  theme(
    panel.background = element_rect(fill = "grey95", color = NA),
    panel.grid.major = element_line(color = "white", size = 1),
    panel.grid.minor = element_line(color = "white", size = 0.4),
    strip.background = element_rect(fill = "grey85", color = NA),
    strip.text = element_text(face = "bold", size = 13),
    plot.title = element_text(face = "bold", size = 14),
    plot.subtitle = element_text(size = 10, face = "italic"),
    axis.text = element_text(size = 10, face = "bold"),
    axis.text.x = element_text(angle = 30, hjust = 1),
    axis.title = element_text(face = "bold", size = 11),
    legend.position = "none"
  )
```

Assignment 3 Task 1

AI sources and ChatGPT Prompts used

1)
Prompt:
“can u help me label lines directly instead of using a legend in a ggplot line chart?

```
ggplot(data = df_filtered, aes(x = Time, y = Value, color = Ratio)) +
  geom_line(size = 0.6) +
  facet_wrap(~Location, ncol = 2) +
  scale_color_manual(
    values = c("Child Dependency" = "#1f77b4", "Old-Age Dependency" = "#d62728")
  ) +
  labs(
    title = "Dependency Trends by Income Group (2000–2024)",
    subtitle = "..."
  )”
```

Line used:

```
geom_text(data = df_labels, aes(label = Ratio, vjust = vjust, hjust = hjust), size = 3, show.legend = FALSE)
```

2)
Prompt:
“The line labels at the end of my line plot are cut off how do I stop that

```
geom_text(data = df_labels, aes(label = Ratio), size = 3)”
```

Line used:

```
“geom_text(data = df_labels, aes(label = Ratio, vjust = -0.5, hjust = 0.5), size = 3, show.legend = FALSE)”
```

Internet Sources

- For preventing cut off geom_text () labels at the plot edges

https://stackoverflow.com/questions/12160908/ggplot2-geom-text-label-cut-off

Assignment 4 Task 1

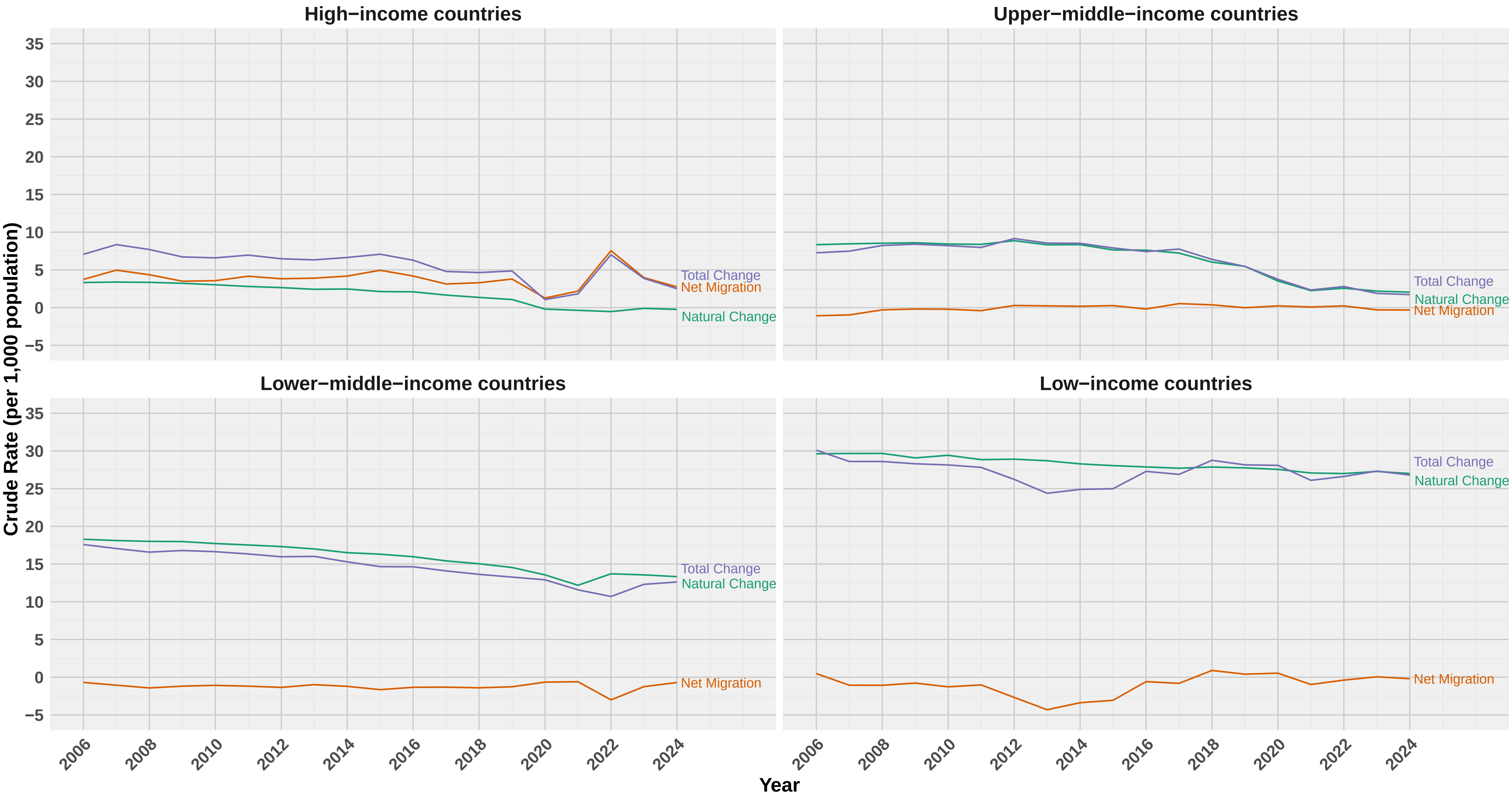
Statement of Purpose

This line chart shows how natural population change, net migration, and the total population change have shifted from 2006 to 2024 in four income-based country groupings. My goal is to help viewers understand the different drivers of population growth. In lower-income countries, natural increase (births minus deaths) remains the key source of growth, while in high-income countries, migration plays a more significant role. The visualization also reveals how migration can offset declining birth rates in some regions, which is important for policymakers thinking about labor force planning and immigration policy.

Trends in Population Change by Income Group (2006−2024)

In low−income countries, population growth is fueled by births. In high−income countries, it's driven by migration. Between 2006 and 2024, lower−income countries experienced strong natural population growth, often above 25 per 1,000 people, despite steady losses from migration. Meanwhile, high−income countries saw declining birth rates and even negative natural change, with migration playing an increasingly critical role in keeping their populations growing. These differences reveal how demographic trends shift depending on income level.

Source: UN Population Division, World Population Prospects (2024).



Assignment 4 Task 1

2025-04-23

R Markdown

```r
library(dplyr)
library(tidyr)
library(ggplot2)
library(readr)
library(scales)

crudedata <- read.csv("/Users/talakammar/Library/Mobile Documents/com~apple~CloudDocs/Spring 2025/Data Visualizat
ion/UNPopulationData_CrudeRates.csv")

df_clean <- crudedata %>%
  filter(IndicatorName %in% c("Crude rate of natural change of population", "Crude rate of net migration"),
         Location %in% c("Low-income countries", "Lower-middle-income countries",
                         "Upper-middle-income countries", "High-income countries"),
         EstimateType == "Model-based Estimates",
         Variant == "Median",
         Time >= 2000, Time <= 2024) %>%
  mutate(Component = ifelse(IndicatorShortName == "Crude rate of natural change of population",
                            "Natural Change", "Net Migration")) %>%
  select(Time, Location, Component, Value)

df_total <- df_clean %>%
  pivot_wider(names_from = Component, values_from = Value) %>%
  mutate(`Total Change` = `Natural Change` + `Net Migration`) %>%
  pivot_longer(cols = c("Natural Change", "Net Migration", "Total Change"),
               names_to = "Component", values_to = "Value") %>%
  filter(Time >= 2006) %>%
  mutate(Location = factor(Location, levels = c("High-income countries", "Upper-middle-income countries",
                                                "Lower-middle-income countries", "Low-income countries")))

df_labels <- df_total %>%
  filter(Time == 2024) %>%
  mutate(vjust = case_when(
    Component == "Total Change" ~ -0.9,
    Component == "Natural Change" ~ 1.2,
    Component == "Net Migration" ~ 0.5
  ))

# Plot
ggplot(data = df_total, aes(x = Time, y = Value, color = Component)) +
  geom_line(size = 0.5) +
  geom_text(data = df_labels, aes(label = Component, vjust = vjust),
            hjust = -0.05, size = 3, show.legend = FALSE) +
  facet_wrap(~Location, ncol = 2) +
  scale_color_manual(values = c("Natural Change" = "#1b9e77",
                                "Net Migration" = "#d95f02",
                                "Total Change" = "#7570b3")) +
  scale_y_continuous(limits = c(-5, 35), breaks = seq(-5, 35, 5)) +
  scale_x_continuous(breaks = seq(2006, 2024, 2), limits = c(2006, 2026)) +
  labs(
    title = "Trends in Population Change by Income Group (2006–2024)",
    subtitle =
      "In low-income countries, population growth is fueled by births. In high-income countries, it’s driven by m
igration. Between 2006 and 2024, lower-income countries experienced strong natural population growth,
often above 25 per 1,000 people, despite steady losses from migration. Meanwhile, high-income countries saw decli
ning birth rates and even negative natural change, with migration playing an increasingly
critical role in keeping their populations growing. These differences reveal how demographic trends shift dependi
ng on income level.

Source: UN Population Division, World Population Prospects (2024).",
    x = "Year",
    y = "Crude Rate (per 1,000 population)"
  ) +
  theme_minimal(base_size = 10) +
  theme(
    panel.grid.major = element_line(color = "gray80", size = 0.4),
    panel.grid.minor = element_line(color = "gray90", size = 0.2),
    panel.background = element_rect(fill = "#f0f0f0", color = NA),
    strip.text = element_text(face = "bold", size = 13),
    plot.title = element_text(face = "bold", size = 12),
    plot.subtitle = element_text(size = 10, face = "italic"),
    axis.title.x = element_text(size = 13, face = "bold"),
    axis.title.y = element_text(size = 13, face = "bold"),
    axis.text = element_text(size = 11, face = "bold"),
    axis.text.x = element_text(angle = 45, hjust = 1),
    legend.position = "none"
  )
```

Assignment 4 Task 1

AI sources and ChatGPT Prompts used

1)
Prompt:

“Im trying to show all three lines ‘Natural Change,’ ‘Net Migration,’ ‘Total Change’ in one plot and make multiple income group facets. Did I pivot correctly? Code is below:

```
df_total <- df_clean %>%
  pivot_wider(names_from = Component, values_from = Value) %>%
  mutate(`Total Change` = `Natural Change` + `Net Migration`) %>%
  pivot_longer(cols = c("Natural Change", "Net Migration", "Total Change"),
               names_to = "Component", values_to = "Value")
```

Line used:

```
mutate(`Total Change` = `Natural Change` + `Net Migration`)
```

2)
Prompt:
“I want to directly label my lines on the far right (2024) instead of using a legend. Labels overlapping sometimes. Code below:

```
geom_text(data = df_labels, aes(label = Component), size = 3)”
```

Line used:

```
geom_text(data = df_labels, aes(label = Component, vjust = vjust), hjust = -0.05, size = 3,
show.legend = FALSE)

mutate(vjust = case_when(
  Component == "Total Change" ~ -0.9,
  Component == "Natural Change" ~ 1.2,
  Component == "Net Migration" ~ 0.5
))”
```

3)
Prompt:
"my background panel is too plain. how do I give it a color but while keeping grid lines clear and have the lines layered over it?”

Line used:

```
panel.background = element_rect(fill = "#f0f0f0", color = NA)
```

4)
Prompt:
“I want my x axis to extend a bit beyond the last data point so the labels I add using geom_text() aren’t cut off”

Line used:
scale_x_continuous(breaks = seq(2006, 2024, 2), limits = c(2006, 2026))

Internet Sources

- For Fixing labels in geom_text() being cut off at the plot edges

https://stackoverflow.com/questions/12160908/ggplot2-geom-text-label-cut-off

- For computing and re-pivot multiple variables in ggplot-ready format. Validating use of pivot_wider() -> mutate() -> pivot_longer() chain when plotting derived variables like “Total Change.”

https://stackoverflow.com/questions/72052489/how-to-sum-values-and-pivot-data-for-ggplot-in-r

- For guidance on layering a colored background with grid lines intact for facets:

https://stackoverflow.com/questions/31852047/custom-background-color-in-ggplot2

Assignment 4 Task 2

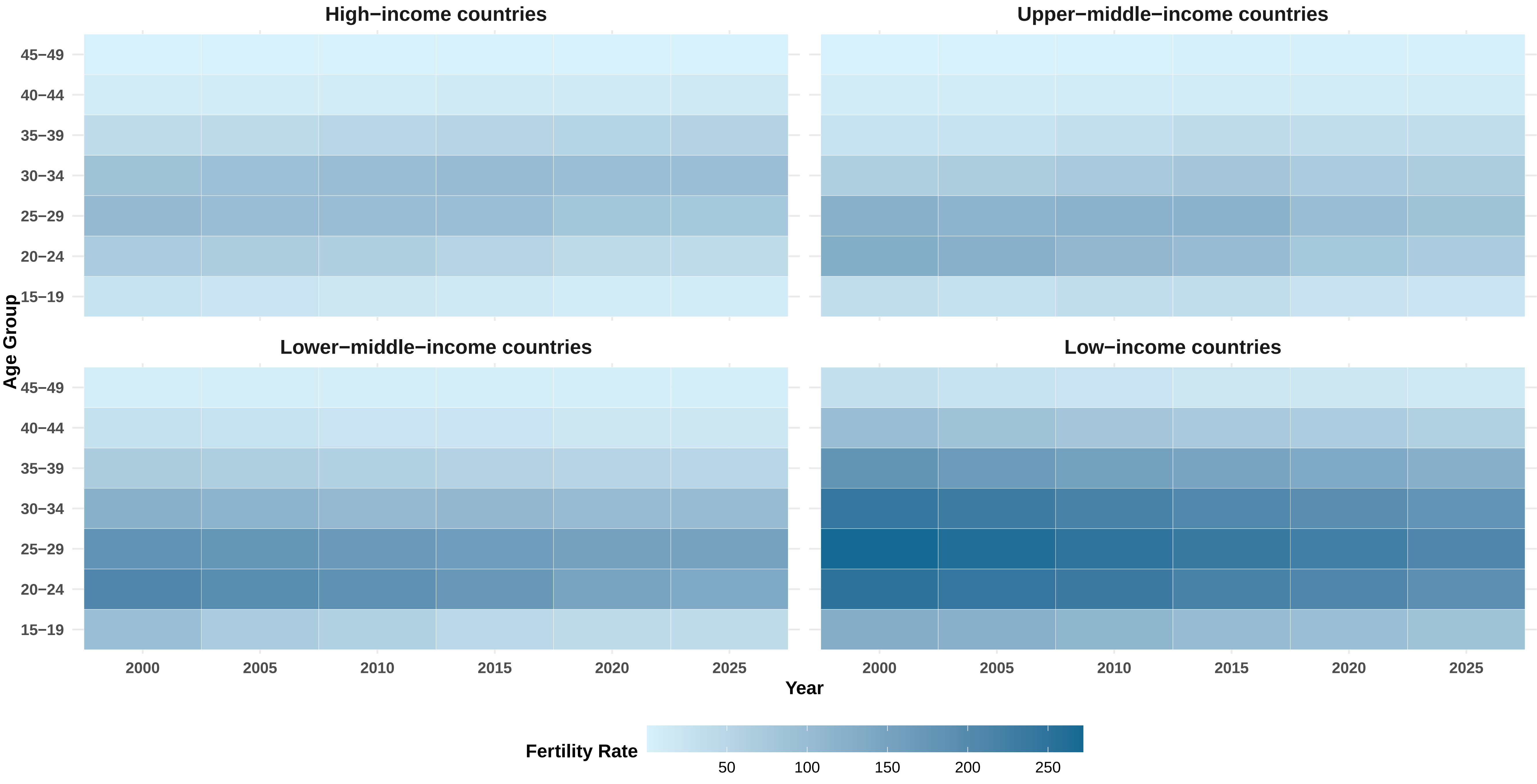
Statement of Purpose

The following heatmap presents fertility rates by age group across income groups from 2000 to 2025. By organizing the data by age and year, the chart shows not only when fertility tends to peak in each income group, but also how the overall intensity and timing of childbearing have shifted over time. I created this visualization to emphasize the contrast between early fertility in low-income countries and delayed or lower fertility in high-income ones. The chart helps show how access to education, healthcare, and family planning varies across regions and influences these patterns.

Fertility Patterns by Age and Income Group

Across all income groups, fertility peaks between ages 20−24, but the timing and intensity vary widely. Low−income countries show high fertility at younger ages, while high−income countries experience lower rates and delayed childbearing. Overall, fertility has declined across age groups over time, especially among adolescents−reflecting shifts in access to education, health services, and family planning.

Source: UN Population Division, World Population Prospects (2024).



Assignment 4 Task 2

2025-04-23

R Markdown

```r
library(dplyr)
library(tidyr)
library(ggplot2)
library(readr)
library(scales)

fertility_data <- read.csv("/Users/talakammar/Library/Mobile Documents/com~apple~CloudDocs/Spring 2025/Data Visualization/UNPopulationData_Fertility.csv")

filtered_data <- fertility_data %>%
  filter(Time %in% c(2000, 2005, 2010, 2015, 2020, 2025),
         Age %in% c("15-19", "20-24", "25-29", "30-34", "35-39", "40-44", "45-49"),
         Location %in% c("Low-income countries", "Lower-middle-income countries",
                         "Upper-middle-income countries", "High-income countries"),
         EstimateType == "Model-based Estimates",
         Variant == "Median") %>%
  mutate(
    Age = factor(Age, levels = c("15-19", "20-24", "25-29", "30-34", "35-39", "40-44", "45-49")),
    Time = as.factor(Time),
    Location = factor(Location, levels = c("High-income countries", "Upper-middle-income countries",
                                           "Lower-middle-income countries", "Low-income countries"))
  )

mean_value <- mean(filtered_data$Value, na.rm = TRUE)

ggplot(data = filtered_data,
       mapping = aes(x = Time, y = Age)) +
  geom_tile(aes(fill = Value), color = "white") +
  scale_fill_gradient(
    low = "#d8f0f9",
    high = "#146993",
    name = "Fertility Rate",
    breaks = seq(0, 300, 50)
  ) +
  facet_wrap(~Location, ncol = 2) +
  labs(
    title = "Fertility Patterns by Age and Income Group",
    subtitle = "
Across all income groups, fertility peaks between ages 20-24, but the timing and intensity vary widely. Low-income countries show high fertility at younger ages, while high-income countries
experience lower rates and delayed childbearing. Overall, fertility has declined across age gr
oups over time, especially among adolescents—reflecting shifts in access to education, health
services, and
family planning.\n\nSource: UN Population Division, World Population Prospects (2024).",
    x = "Year",
    y = "Age Group"
  ) +
  theme_minimal(base_size = 12) +
  theme(
    strip.text = element_text(face = "bold", size = 13),  # standardized facet size
    plot.title = element_text(face = "bold", size = 14),
    plot.subtitle = element_text(size = 10, face = "italic"),
    axis.title = element_text(face = "bold", size = 12),
    axis.text = element_text(size = 10, face = "bold"),
    legend.position = "bottom",
    legend.key.width = unit(2, "cm"),
    legend.text = element_text(size = 10),
    legend.title = element_text(face = "bold"),
    panel.background = element_blank()
  )
```

Assignment 4 Task 2

AI sources and ChatGPT Prompts used

1)
Prompt:
“How can I visually show fertility trends across both age groups and time periods using a heatmap in ggplot2?”

Line used:

```
geom_tile(aes(fill = Value), color = "white")
```

2)
Prompt:
“I want to highlight multiple variables in a heatmap, including age, year, and income group. What’s the best way to structure my data in R?”

Line used:

```
facet_wrap(~Location, ncol = 2)
```

3)
Prompt:
“How do I reorder a factor in R so that ggplot displays the ages in a logical order from youngest to oldest?”

Line used:

```
Age = factor(Age, levels = c("15-19", "20-24", "25-29", "30-34", "35-39", "40-44", "45-49"))
```

4)
Prompt:
“What’s a colorblind-friendly gradient I can use to represent intensity in a heatmap without using red/green?”

Line used:

```
scale_fill_gradient(low = "#d8f0f9", high = "#146993", name = "Fertility Rate")
```

Internet Sources:

- Used for reording factor levels for display in ggplot2:
https://stackoverflow.com/questions/5208679/order-bars-in-ggplot2-bar-graph

- For help troubleshooting heatmap tile alignment, spacing, white gaps in ggplot2:
https://stackoverflow.com/questions/28482150/how-to-adjust-spacing-in-ggplot2-heatmap

- Referenced to properly choose and set up a visually appropriate and accessible fill color scale:

https://ggplot2.tidyverse.org/reference/scale_gradient.html

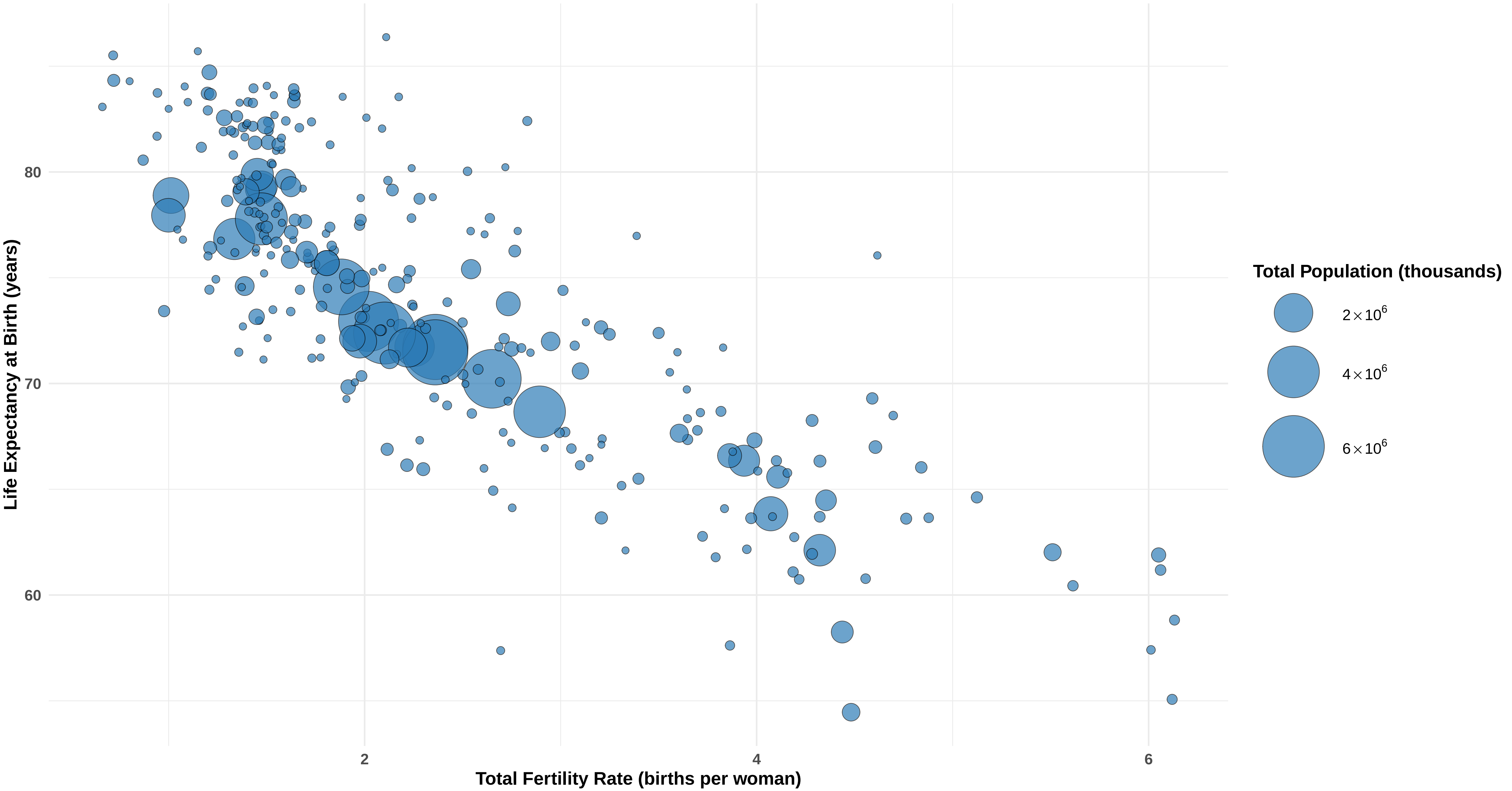
Assignment 5 Task 1

Statement of Purpose

This visualization examines the relationship between fertility rates and life expectancy across countries in 2023. Each bubble represents a country, with bubble size corresponding to total population. The plot highlights how countries with lower fertility rates tend to have higher life expectancies, reflecting advanced stages of demographic transition. In contrast, countries with higher fertility rates often experience shorter life expectancies, indicating earlier stages of development. By capturing both population size and demographic indicators, this plot demonstrates how demographic pressures vary widely depending on where a country is in its development journey.

Global Life Expectancy and Fertility Trends (2023)

Across all regions, lower fertility is associated with longer life expectancy. High−fertility countries face younger populations, while low−fertility countries are aging rapidly.
These trends reflect different demographic transitions shaping global development.



Source: United Nations, World Population Prospects 2024

Assignment 5 Task 1

2025-05-02

```
library(ggplot2)
library(readxl)
library(dplyr)
library(scales)

population_data <- read_excel("/Users/talakammar/Library/Mobile Documents/com~apple~CloudDocs/
Spring 2025/Data Visualization/WPP2024_GEN_F01_DEMOGRAPHIC_INDICATORS_FULL (1).xlsx",range = "
A17:BM22000")

df_plot <- population_data %>%
  filter(
    Year == 2023,
    !`Region, subregion, country or area *` %in% c(
      "World", "High-income countries", "Low-income countries",
      "Lower-middle-income countries", "Upper-middle-income countries"
    )
  ) %>%
  mutate(
    `Total Population, as of 1 July (thousands)` = as.numeric(`Total Population, as of 1 July
(thousands)`),
    `Total Fertility Rate (live births per woman)` = as.numeric(`Total Fertility Rate (live bi
rths per woman)`),
    `Life Expectancy at Birth, both sexes (years)` = as.numeric(`Life Expectancy at Birth, bot
h sexes (years)`)
  )

scientific_10 <- function(x) {
  parse(text = gsub("e\\+*", " %*% 10^", scientific_format()(x)))
}

ggplot(df_plot, aes(
  x = `Total Fertility Rate (live births per woman)`,
  y = `Life Expectancy at Birth, both sexes (years)`,
  size = `Total Population, as of 1 July (thousands)`
)) +
  geom_point(shape = 21, fill = "#2C7BB6", color = "black", stroke = 0.3, alpha = 0.7) +
  scale_size(
    range = c(2, 20),
    labels = scientific_10
  ) +
  labs(
    title = "Global Life Expectancy and Fertility Trends (2023)",
    subtitle = "
Across all regions, lower fertility is associated with longer life expectancy. High-fertility
countries face younger populations, while low-fertility countries are aging rapidly.
These trends reflect different demographic transitions shaping global development.",
    caption = "Source: United Nations, World Population Prospects 2024",
    x = "Total Fertility Rate (births per woman)",
    y = "Life Expectancy at Birth (years)",
    size = "Total Population (thousands)"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 16, face = "bold"),
    plot.subtitle = element_text(size = 11, face = "italic", margin = margin(b = 10)),
    plot.caption = element_text(size = 9, face = "italic", hjust = 0, margin = margin(t = 10))
,
    axis.title = element_text(size = 12, face = "bold"),
    axis.text = element_text(size = 10, face = "bold"),
    legend.title = element_text(size = 12, face = "bold"),
    legend.text = element_text(size = 10)
  )
```

Assignment 5 Task 1

AI sources and ChatGPT Prompts used

1)
Prompt:
“im trying to do a bubble plot where bubble size = population, x = fertility, y = life expectancy. heres what i have so far:

```
ggplot(df_plot, aes(
  x = `Total Fertility Rate (live births per woman)`,
  y = `Life Expectancy at Birth, both sexes (years)`,
  size = `Total Population, as of 1 July (thousands)`
)) +
  geom_point()”
```

Line used:

```
geom_point(shape = 21, fill = "#2C7BB6", color = "black", stroke = 0.3, alpha = 0.7)
```

2)
Prompt:
“how can i make my bubble sizes in powers of 10 instead of showing numbers with e”

Line used:

```
scientific_10 <- function(x) {
  parse(text = gsub("e\\+*", " %*% 10^", scientific_format()(x)))
}

ggplot(df_plot, aes(
  x = `Total Fertility Rate (live births per woman)`,
  y = `Life Expectancy at Birth, both sexes (years)`,
  size = `Total Population, as of 1 July (thousands)`
)) +
  geom_point(shape = 21, fill = "#2C7BB6", color = "black", stroke = 0.3, alpha = 0.7) +
  scale_size(
    range = c(2, 20),
    labels = scientific_10
  )”
```

3)
Prompt:
“i wanna plot the individual countries in the various income groups. Should i use filter() and %in% or smth else?

Line used:

```
filter(
  Year == 2023,
  !`Region, subregion, country or area *` %in% c(
    "World", "High-income countries", "Low-income countries",
    "Lower-middle-income countries", "Upper-middle-income countries"
  )
)
```

4)
Prompt:
“some of my numeric columns are showing up as char. convert them without rewriting everything. Here they are Total Population, as of 1 July (thousands), Total Fertility Rate (live births per woman), and Life Expectancy at Birth, both sexes (years)”

Line used:

```
mutate(
    `Total Population, as of 1 July (thousands)` = as.numeric(`Total Population, as of 1 July (thousands)`),
    `Total Fertility Rate (live births per woman)` = as.numeric(`Total Fertility Rate (live births per woman)`),
    `Life Expectancy at Birth, both sexes (years)` = as.numeric(`Life Expectancy at Birth, both sexes (years)`)
  )
```

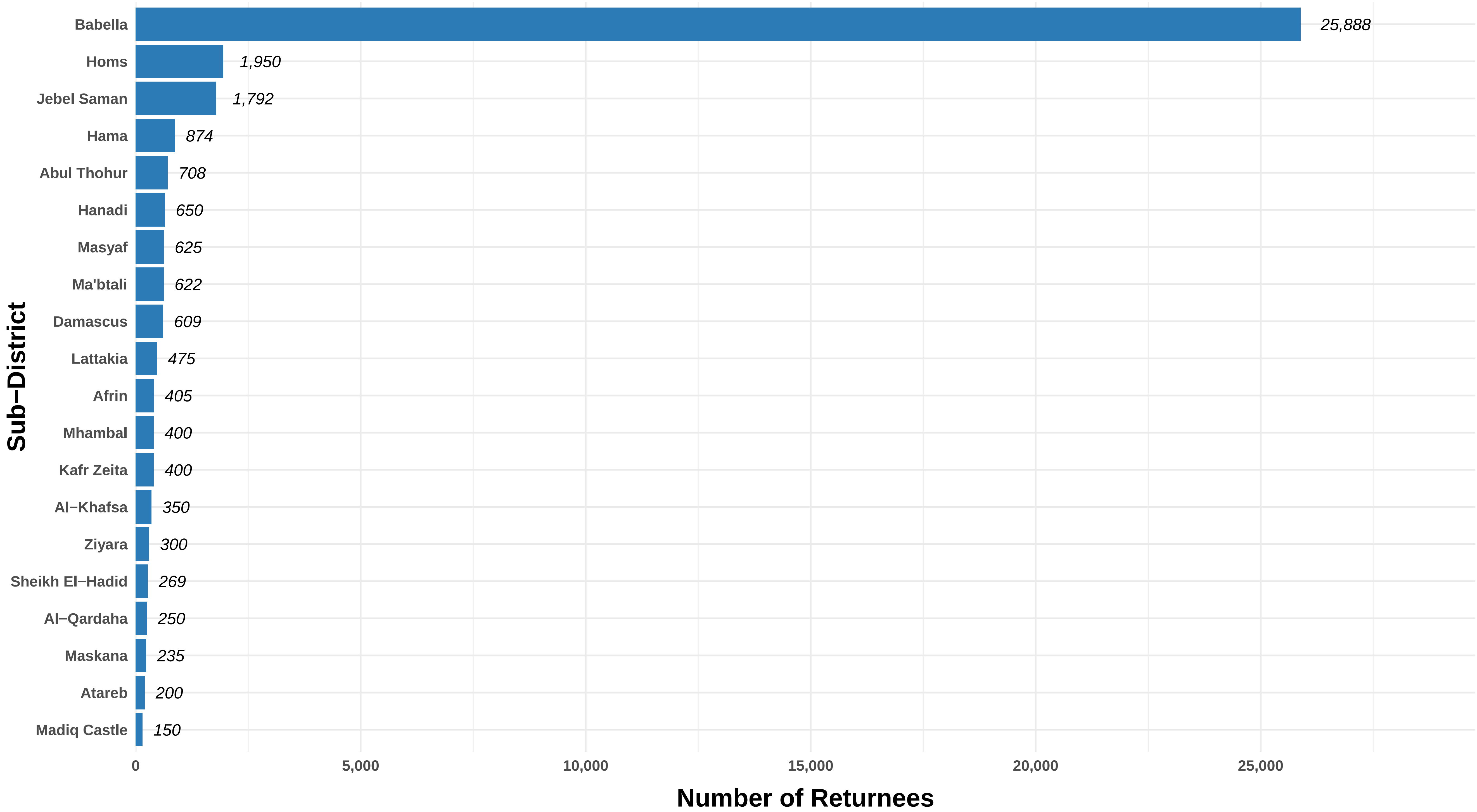
Assignment 5 Task 2

Statement of Purpose

This visualization shows the top 20 sub-districts in Syria with the highest number of returnees as of November 2024, only weeks before the regime fell on December 8. Each bar represents a sub-district, with the length of the bar corresponding to the total number of individuals who returned from internal displacement. Babella stands out as a dramatic outlier, receiving far more returnees than any of the others. What's interesting is that this return movement was already happening right before a major political shift. While returns were happening across the country, the fact that they were concentrated in certain areas might point to zones that were seen as safer or more stable. The chart gives a clear look at how resettlement was taking shape in real time offering some clues about where recovery efforts or support might be most needed moving forward. It could be helpful to monitor changes in returnee numbers over time across these same sub-districts, especially post-regime fall. Tracking trends could reveal whether return patterns stabilize, reverse, or accelerate which could help inform where post war reconstruction efforts and aid should be focused.

Top 20 Sub−Districts in Syria by Number of Returnees Since November 2024

Babella stands out with a significantly higher returnee count than all other sub−districts.



Source: United Nations, International Organization for Migration (IOM). Displacement Tracking Matrix (DTM) − Syria: Baseline Assessment Round 1, February 2025.

Assignment 5 Task 2

2025-05-02

```r
library(ggplot2)
library(readxl)
library(dplyr)
library(scales)


syria_data <- read_excel("/Users/talakammar/Library/Mobile Documents/com~apple~CloudDocs/Spring 2025/Data Visualization/dtm-syria-baseline-assessment-round-1-feb-25.xlsx",
  sheet = "BaselineAssessment_Feb2025",
  range = "A4:YM8511"
)

colnames(syria_data) <- make.names(colnames(syria_data))

names(syria_data)[grepl("Governorate", names(syria_data))]
names(syria_data)[grepl("District", names(syria_data))]
names(syria_data)[grepl("Sub.district", names(syria_data))]
names(syria_data)[grepl("RETURNEES.IND.in.November.2024", names(syria_data))]

syria_data_filtered <- syria_data %>%
  select(
    Governorate = Governorate,
    District = District,
    SubDistrict = Sub.district,
    Returnees = RETURNEES.IND.in.November.2024
  ) %>%
  filter(!is.na(Returnees)) %>%
  group_by(Governorate, District, SubDistrict) %>%
  summarise(TotalReturnees = sum(Returnees, na.rm = TRUE), .groups = "drop") %>%
  arrange(desc(TotalReturnees))

top_returnees <- syria_data_filtered %>% slice_max(TotalReturnees, n = 20)

ggplot(top_returnees, aes(x = reorder(SubDistrict, TotalReturnees), y = TotalReturnees)) +
  geom_col(fill = "#2C7BB6") +
  geom_text(aes(label = scales::comma(TotalReturnees)),
            hjust = -0.4, size = 4, fontface = "italic") +
  coord_flip() +
  labs(
    title = "Top 20 Sub-Districts in Syria by Number of Returnees Since November 2024",
    subtitle = "Babella stands out with a significantly higher returnee count than all other sub-districts.",
    caption = "Source: United Nations, International Organization for Migration (IOM). Displacement Tracking Matrix (DTM) - Syria: Baseline Assessment Round 1, February 2025.",
    x = "Sub-District",
    y = "Number of Returnees"
  ) +
  scale_y_continuous(
    breaks = seq(0, max(top_returnees$TotalReturnees) + 2000, by = 5000),
    labels = label_comma(),
    expand = expansion(mult = c(0, 0.15))
  ) +
  theme_minimal(base_size = 12) +
  theme(
    plot.title = element_text(face = "bold", size = 22),
    plot.subtitle = element_text(face = "italic", size = 13),
    plot.caption = element_text(size = 10, face = "italic", hjust = 0, vjust = -0.8, margin =
margin(t = 10)),
    axis.title = element_text(size = 12, face = "bold"),
    axis.title.y = element_text(size = 17, face = "bold", vjust = -2),
    axis.title.x = element_text(size = 17, face = "bold", vjust = -1),
    axis.text.x = element_text(face = "bold"),
    axis.text.y = element_text(face = "bold")
  )
```

Assignment 5 Task 2

AI sources and ChatGPT Prompts used

1)
Prompt:
“im plotting returnees by sub district and wanna add the actual numbers directly on the bars. i want them italic and a bit away from the bars so they dont look all cramped. heres what i got so far but the numbers arent showing right, help pls”

```
ggplot(top_returnees, aes(x = reorder(SubDistrict, TotalReturnees), y = TotalReturnees)) +
  geom_col()
```

Line used:

```
geom_text(aes(label = scales::comma(TotalReturnees)),
          hjust = -0.4, size = 3.5, fontface = "italic")
```

2)
Prompt:
“only some of the y axis ticks are showing like 10k andn 20k but i want all the gridlines to show with labels. is there a way to just manually set them? current y scale isnt doin that”

```
scale_y_continuous(labels = label_comma())
```

Line used:

```
scale_y_continuous(
  breaks = seq(0, max(top_returnees$TotalReturnees) + 2000, by = 5000),
  labels = label_comma(),
  expand = expansion(mult = c(0, 0.15))
)
```

3)
Prompt:
“i thought theme_minimal() removes backgrounds but im still gettin this random grey box behind the plot?? i just want it clean. what do i remove?”

Fix used:
Just kept theme_minimal() and made sure I didnt accidentally add panel.background() again anywhere else. didnt manually set any fill either so it stayed clean.

4)
Prompt:
“whats the cleanest way to keep the default gridlines but also have small ticks for the smaller intervals like every 5k? rn theyre super spaced out and looks empty”

Line used:
Same thing as in #2, I just used seq() in breaks to add a tick every 5000 manually.

5)
Prompt:
“is there any downside to using reorder(SubDistrict, TotalReturnees) directly in aes() if im already arranging the data with slice_max() before plotting? I dont want double sorting to mess w factor levels”

```
ggplot(top_returnees, aes(x = reorder(SubDistrict, TotalReturnees), y = TotalReturnees))
```

Line used:
Still used reorder() inside aes() since felt cleaner for labeling on the flipped y-axis

6)
Prompt:
“im not setting a ylim() but I wanna make sure the expand = expansion(mult = c(0, 0.15)) wont hide the tallest bar or mess w alignment help pls”

Line used:

```
scale_y_continuous(
  breaks = seq(0, max(top_returnees$TotalReturnees) + 2000, by = 5000),
  labels = label_comma(),
  expand = expansion(mult = c(0, 0.15))
)
```

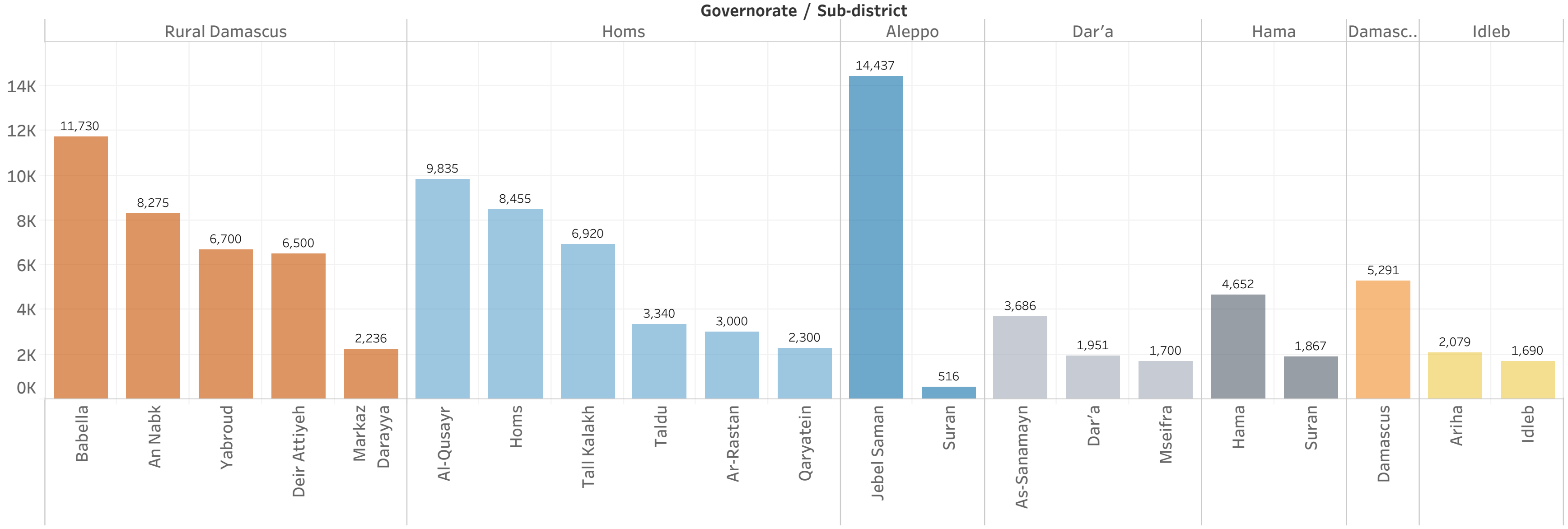
Assignment 6 Task 1

Statement of Purpose

This visualization identifies the 20 sub-districts in Syria that recorded the highest number of returnees to their original places of origin in November 2024, immediately following military operations leading up to a government change in December. The bar chart shown is organized in descending order by governorate and sub-district to emphasize regional concentrations of return activity. The purpose of this chart is to highlight areas where return migration is most significant, providing insight into zones that may require urgent support for reintegration, infrastructure restoration, and services. The high number of returnees in sub-districts such as Jebel Saman and Babella could potentially, but not necessarily, point to a perceived increase in safety and likely to a potential strain on local resources.

Top 20 Sub-Districts by Number of Returnees to their Place of Origin Following Military Operations in Syria in November 2024 Ahead of Government Change in December.

This chart displays the total number of individuals who returned to their original sub-districts in November 2024, following military activity and government change, organized in descending order by governorate and sub-district.



Source: United Nations, Interntional Organization for Migration (IOM). Displacement Tracking Matrix (DTM) - Syria: Baseline Assessment Round 1, February 2024

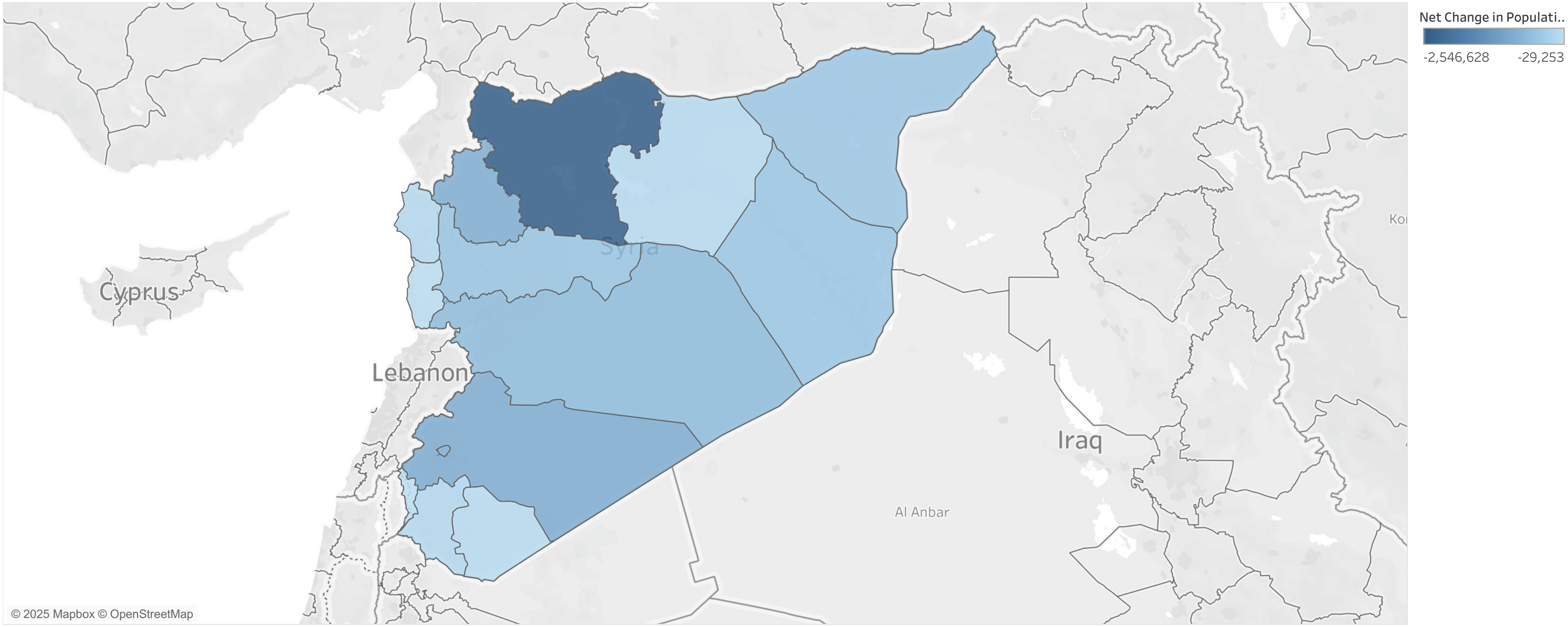
Assignment 6 Task 2

Statement of Purpose

This map displays net population movement by Syrian governorate from 2011 to February 2025. It is calculated as total arrivals minus total departures. The goal is to visualize the cumulative and long-term impact of conflict-driven displacement over time on migration. While every governorate saw an overall loss in population, Aleppo stands out as an outlier, with the most significant drop. The map uses a color gradient to make it easy to compare areas at a glance and understand where movement was most concentrated. The map helps viewers see which parts of the country have been most impacted by displacement over the long term and where recovery efforts might be most urgently needed.

Net Movement by Governorate (2011- February 2025)

This Map shows the net population change per governorate from 2011 to February 2025, calculated as total arrivals minus total departures. While all governorates experienced a net loss, Aleppo stands out as an outlier, experiencing the most significant net change.



© 2025 Mapbox © OpenStreetMap