DATASCI 306, Fall 2024, Final Group Project

Group 2: Connor Tivadar, Alberto Falkon, Alex Matrajt, Jason Arendt, Joseph Painter

Throughout this course, you've dedicated yourself to refining your analytical abilities using R programming language. These skills are highly coveted in today's job market!

Now, for the semester project, you'll apply your learning to craft a compelling `Data Story` that can enrich your portfolio and impress prospective employers. Collaborating with a team (up to 5 members of your choosing), you'll construct a Data Story akin to the example provided here: https://ourworldindata.org/un-population-2024-revision

Data is already in the `data` folder. This data is downloaded from: https://population.un.org/wpp/Download/Standard/MostUsed/

You'll conduct Exploratory Data Analysis (EDA) on the provided data. The provided article already includes 6 diagrams. Show either the line or the map option for these 6 charts. You may ignore the table view. I'm also interested in seeing how each team will expand upon the initial analysis and generate additional 12 insightful charts that includes US and any other region or country that the author did not show. For e.g., one question you may want to answer is; US population is expected to increase to 421 million by 2100. You may want to show how the fertility rate and migration may be contributing to this increase in population.

Deliverable

1. **Requirement-1 (2 pt)** Import the data given in the .xlxs file into two separate dataframes;
   - one dataframe to show data from the `Estimates` tab
   - one dataframe to show data from the `Medium variant` tab

Hint: Some of the steps you may take while importing include:

- skip the first several comment lines in the spread sheet
- Importing the data as text first and then converting the relevant columns to different datatypes in step 2 below.

```
estimates = read_xlsx("data/WPP2024_GEN_F01_DEMOGRAPHIC_INDICATORS_COMPACT.xlsx",
                      sheet = "Estimates", skip = 16, col_types = rep("text", 65))
medium_variant = read_xlsx("data/WPP2024_GEN_F01_DEMOGRAPHIC_INDICATORS_COMPACT.xlsx",
                      sheet = "Medium variant", skip = 16, col_types = rep("text", 65))
```

2. **Requirement-2 (5 pt)**

You should show at least 5 steps you adopt to clean and/or transform the data. Your cleaning should include:

- Renaming column names to make it more readable; removing space, making it lowercase or completely giving a different short name; all are acceptable.
- Removing rows that are irrelevant; look at rows that have Type value as 'Label/Separator'; are those rows required?
- Removing columns that are redundant; For e.g., variant column
- Converting text values to numeric on the columns that need this transformation

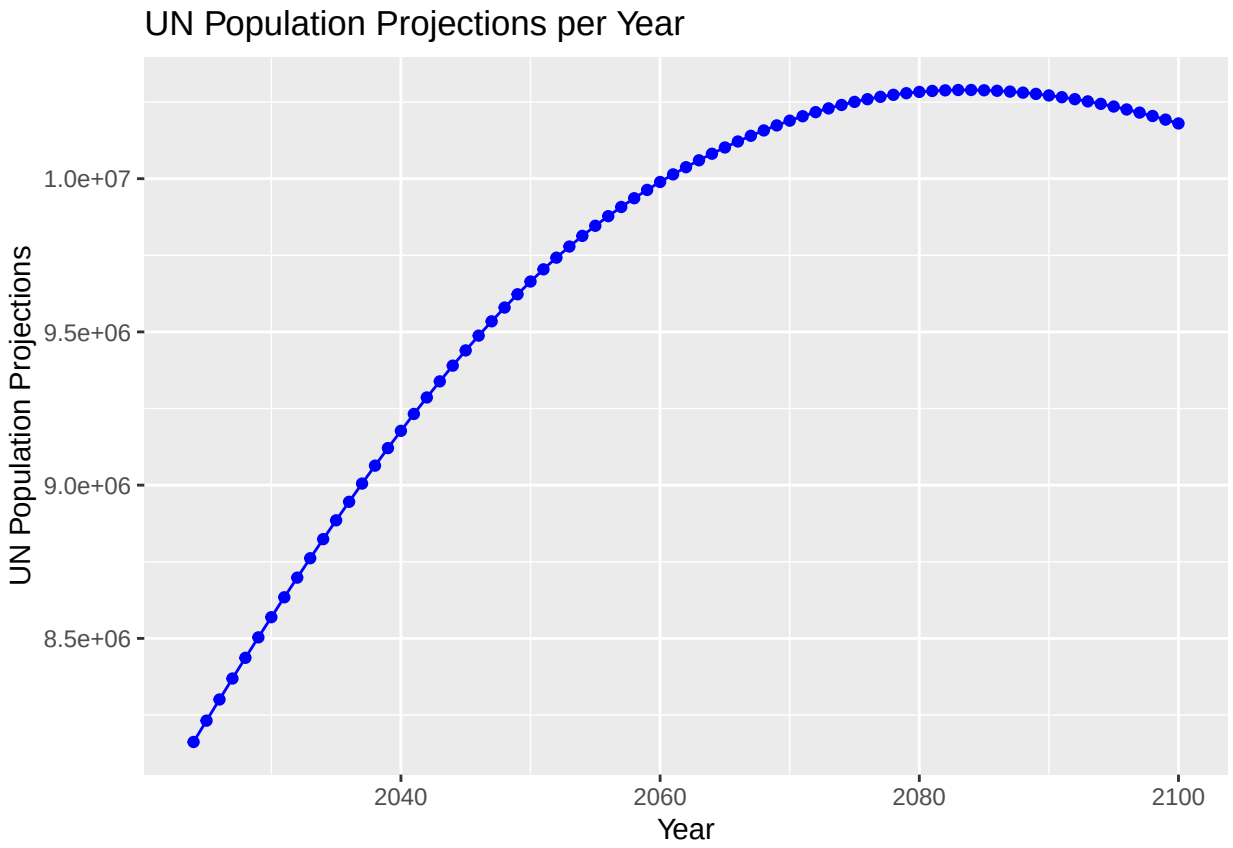
You could also remove the countries/regions that you are not interested in exploring in this step and re-save a smaller file in the same `data` folder, with a different name so that working with it becomes easier going forward.

Explain your reasoning for each clean up step.

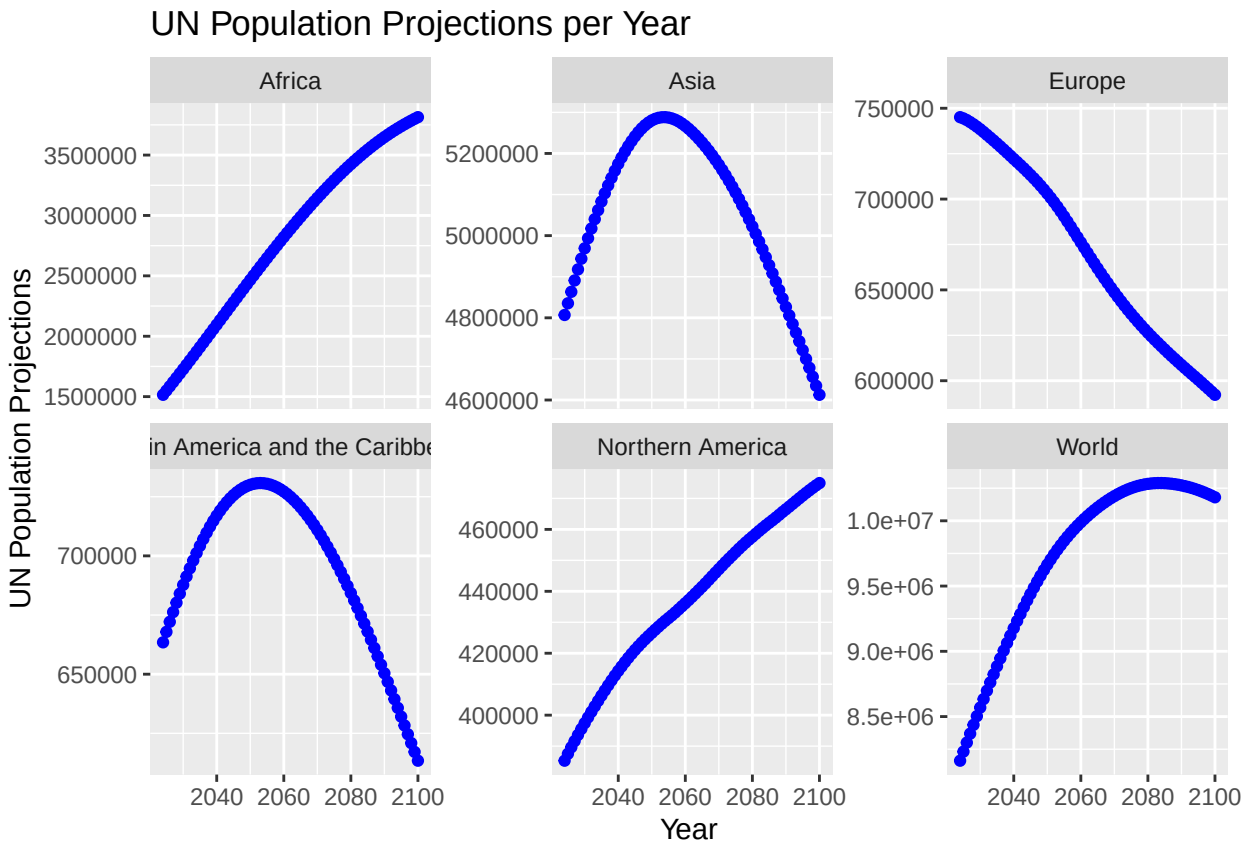
```
clean_and_filter_data <- function(data) {
  data |>
    # Step 1: Normalize column names
    rename_with(~ str_replace_all(tolower(.), "[^a-z0-9]+", "_")) |>
    # Step 2: Remove trailing underscores
    rename_with(~ str_replace(., "_$", "")) |>
    # Step 3: Drop unnecessary columns
    select(-c(variant, notes, iso3_alpha_code, sdmx_code, location_code, iso2_alpha_code, parent_code, i
    # Step 4: Filter out unwanted rows
    filter(type != "Label/Separator") |>
    # Step 5: Convert all columns (except region_subregion_country_or_area and type) to numeric
    mutate(across(
      .cols = !c("region_subregion_country_or_area", "type"),
      .fns = ~ suppressWarnings(as.numeric(gsub("[^0-9.-]", NA_character_, .))),
      .names = "{col}"
    ))
}


estimates_data <- clean_and_filter_data(estimates)
medium_variant_data <- clean_and_filter_data(medium_variant)

#Step 6: Filtering for the columns needed in part 3 and 4
required_medium_variant_data <- medium_variant_data |> select(region_subregion_country_or_area, type, y

required_estimates_data <- estimates_data |> select(region_subregion_country_or_area, type, year, total_
```

3. Requirement-3 (3 pt) Replicate the 6 diagrams shown in the article. Show only the '2024' projection values where ever you have both '2022' and '2024' displayed. Show only the diagrams that are shown on the webpage with default options.

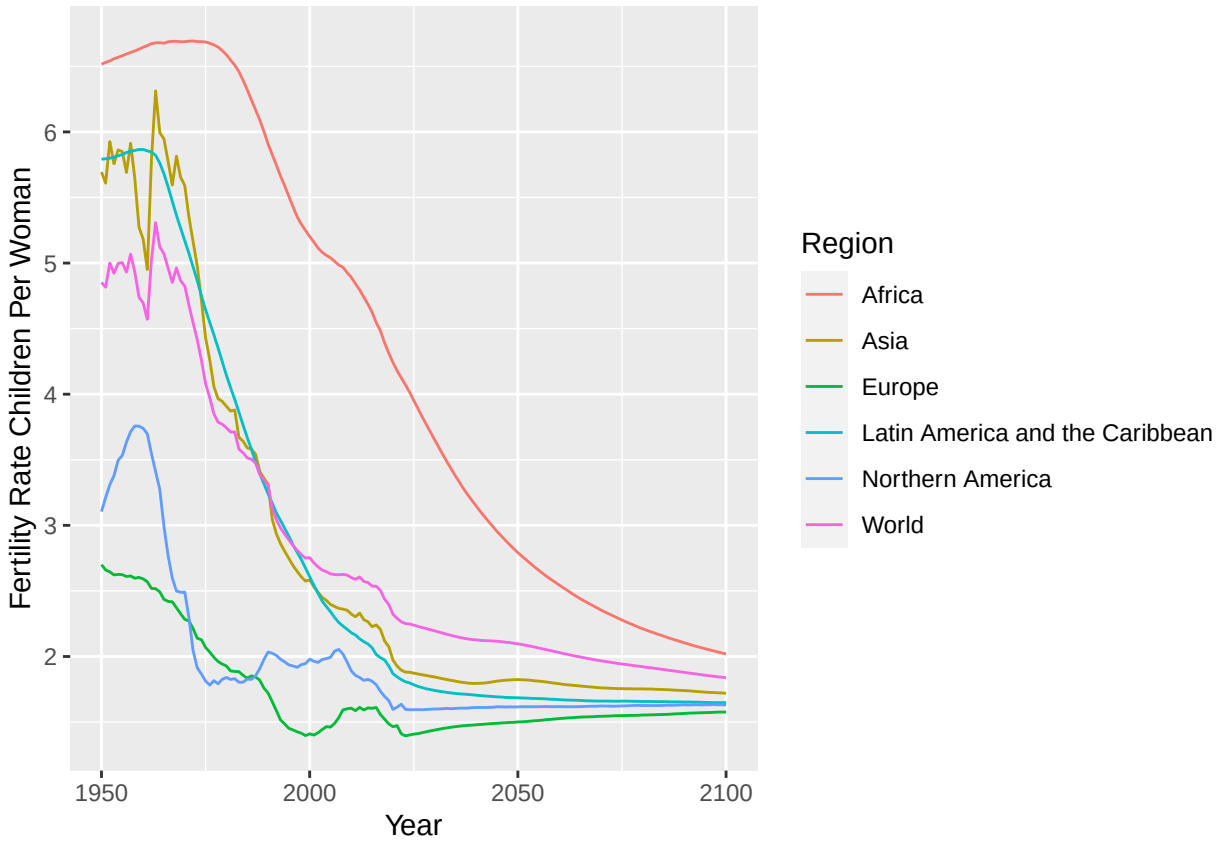
```
required_medium_variant_data |>
  filter(type == "World") |>
  group_by(year) |>
  ggplot(aes(x = year, y = total_population_as_of_1_july_thousands)) +
  geom_line(colour = 'blue') +
  geom_point(colour = 'blue') +
  labs(x = "Year", y = "UN Population Projections", title = "UN Population Projections per Year")
```



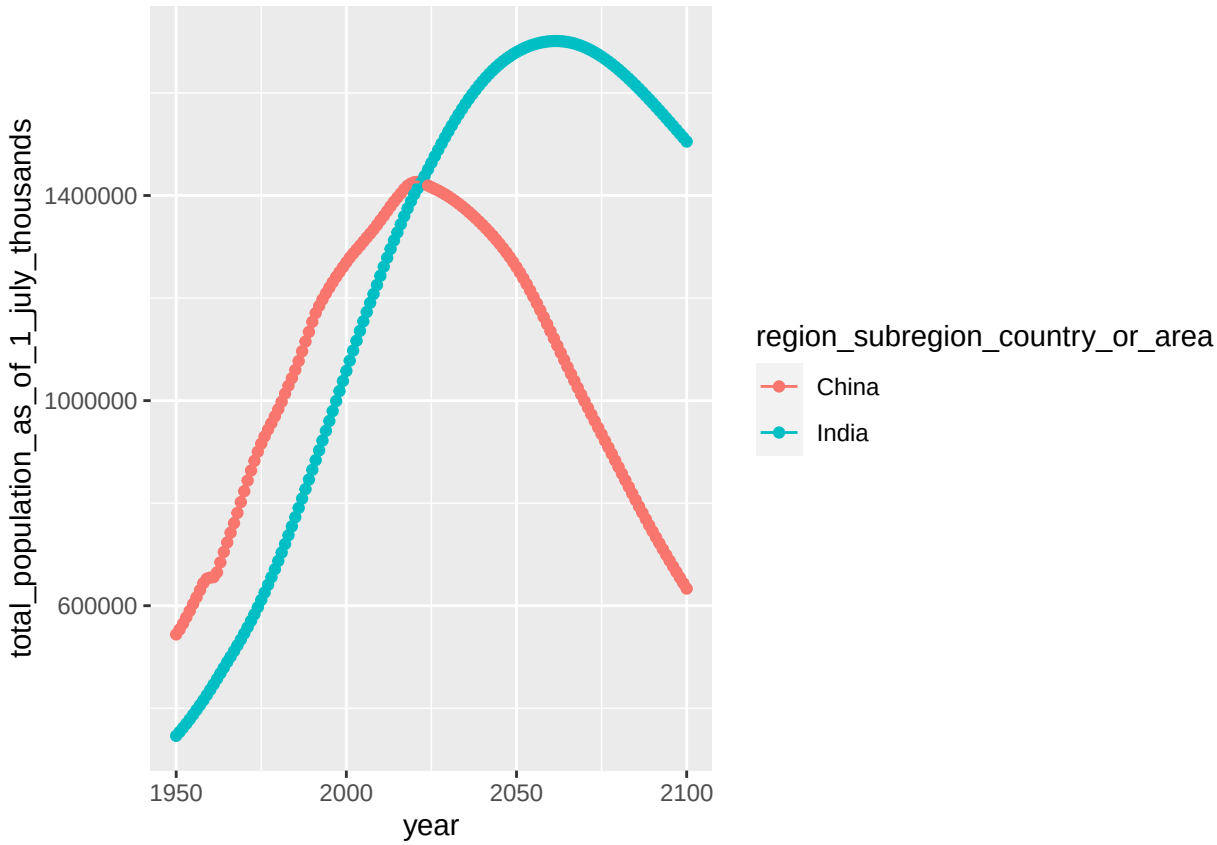

```
required_medium_variant_data |>
  filter(type %in% c("Region", "World")) |>
  filter(region_subregion_country_or_area != "Oceania") |>
  group_by(region_subregion_country_or_area) |>
  ggplot(aes(x = year, y = total_population_as_of_1_july_thousands)) +
  geom_line(colour = 'blue') +
  geom_point(colour = 'blue') +
  facet_wrap(~region_subregion_country_or_area, scales = 'free_y') +
  labs(x = "Year", y = "UN Population Projections", title = "UN Population Projections per Year")
```



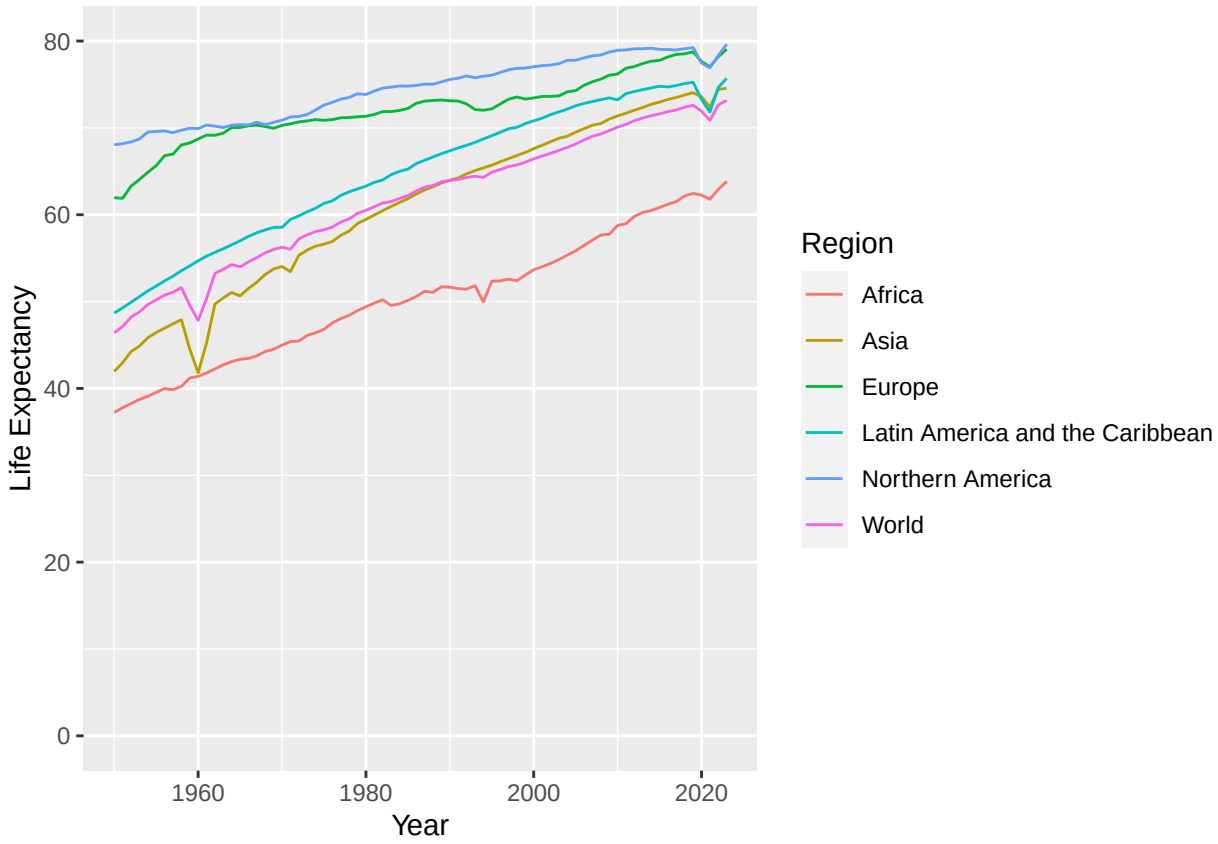

```
required_combine_data <- rbind(required_estimates_data, required_medium_variant_data)
required_combine_data |>
  filter(type %in% c("Region", "World")) |>
  filter(region_subregion_country_or_area != "Oceania") |>
  group_by(year, region_subregion_country_or_area) |>
  ggplot(aes(x = year, y = total_fertility_rate_live_births_per_woman, group = factor(region_subregion_
           geom_line() +
  labs(x = "Year", y = "Fertility Rate Children Per Woman", color = "Region")
```



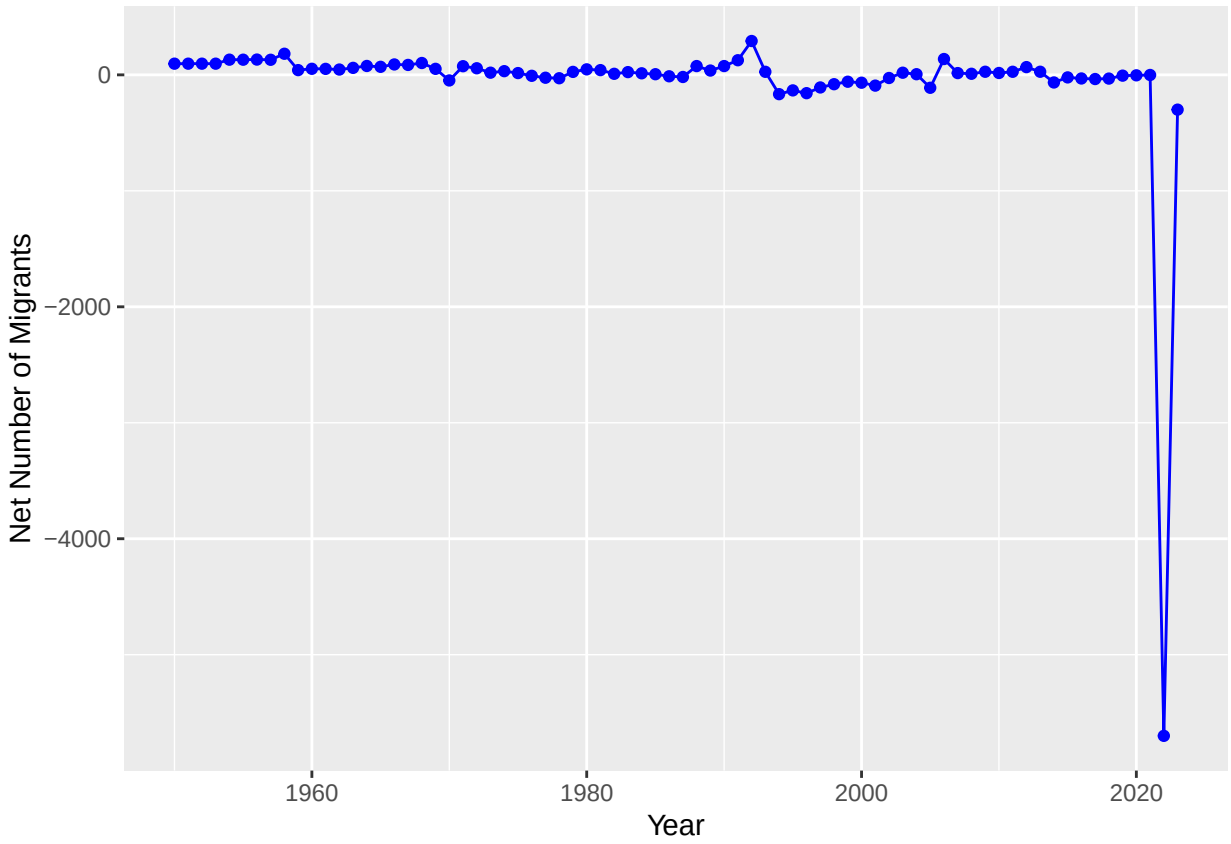

```
required_combine_data |>
  filter(region_subregion_country_or_area %in% c("China", "India")) |>
  group_by(year) |>
  ggplot(aes(x = year, y = total_population_as_of_1_july_thousands, group = factor(region_subregion_cou
           geom_point() +
           geom_line()
```




```
required_estimates_data |>
  filter(type %in% c("Region", "World")) |>
  filter(region_subregion_country_or_area != "Oceania") |>
  group_by(year, region_subregion_country_or_area) |>
  ggplot(aes(x = year, y = life_expectancy_at_birth_both_sexes_years, group = factor(region_subregion_c
           geom_line() +
  labs(x = "Year", y = "Life Expectancy", color = "Region") +
  ylim(0, 80)
```




```
required_estimates_data |>
  filter(region_subregion_country_or_area == "Ukraine") |>
  group_by(year) |>
  ggplot(aes(x = year, y = net_number_of_migrants_thousands)) +
  geom_line(colour = 'blue') +
  geom_point(colour = 'blue') +
  labs(x = "Year", y = "Net Number of Migrants")
```

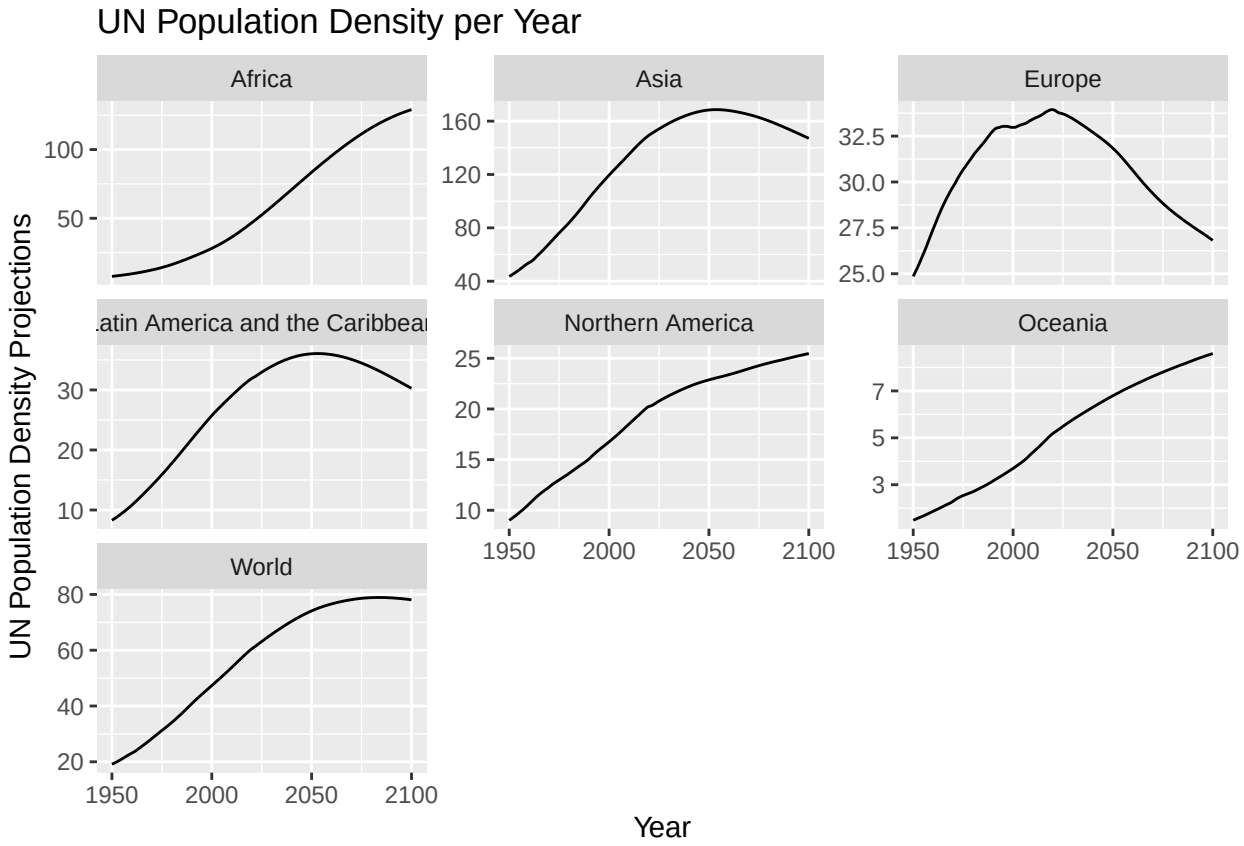


4. Requirement-4 (12 pt)

Select United States related data, and any other country or region(s) of your choosing to perform EDA. Chart at least 12 additional diagrams that may show relationships like correlations, frequencies, trend charts, between various variables with plots of at least 3 different types (line, heatmap, pie, etc.). Every plot should have a title and the x/y axis should have legible labels without any label overlaps for full credit.

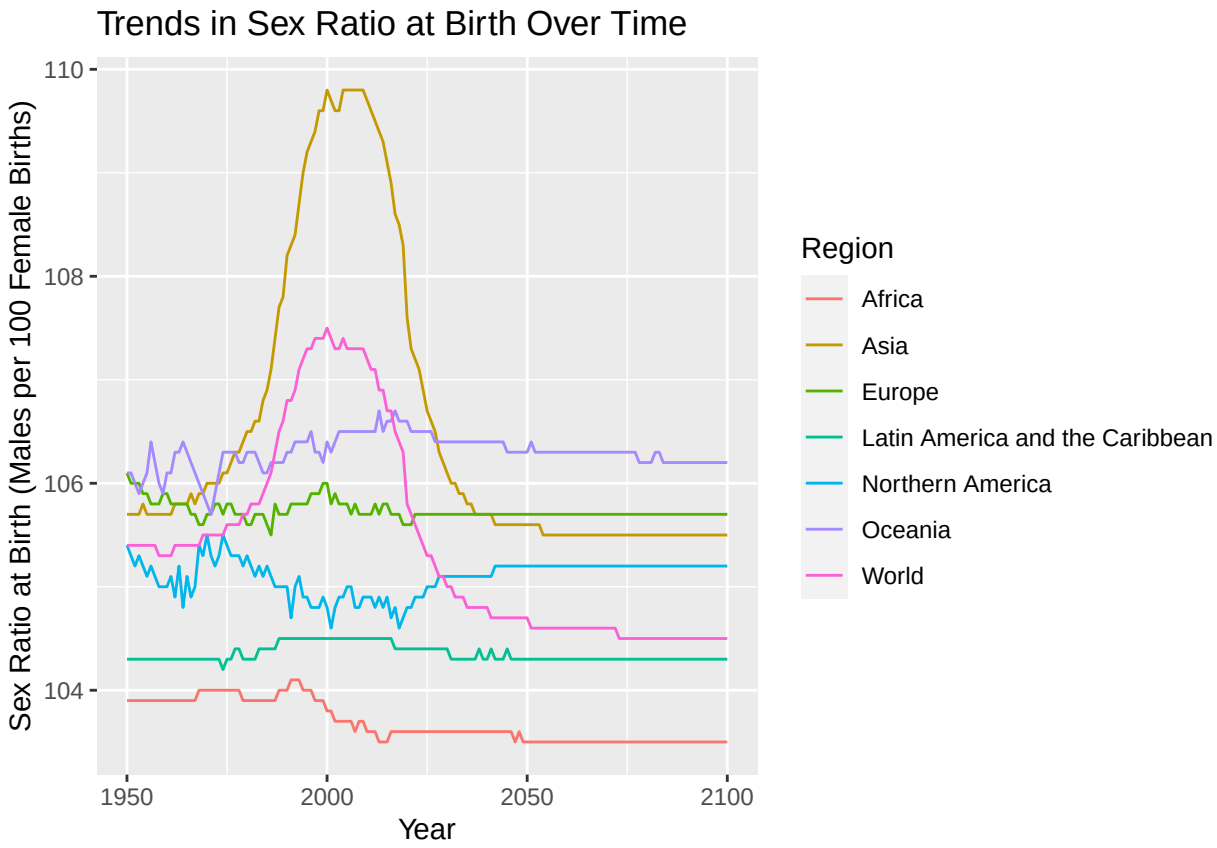
Summarize your interpretations after each chart.

```
required_combine_data |>
  filter(type %in% c("Region", "World")) |>
  group_by(year, region_subregion_country_or_area) |>
  ggplot(aes(x = year, y = population_density_as_of_1_july_persons_per_square_km)) +
  geom_line() +
  facet_wrap(~region_subregion_country_or_area, scales = 'free_y') +
  labs(x = "Year", y = "UN Population Density Projections", title = "UN Population Density per Year")
```



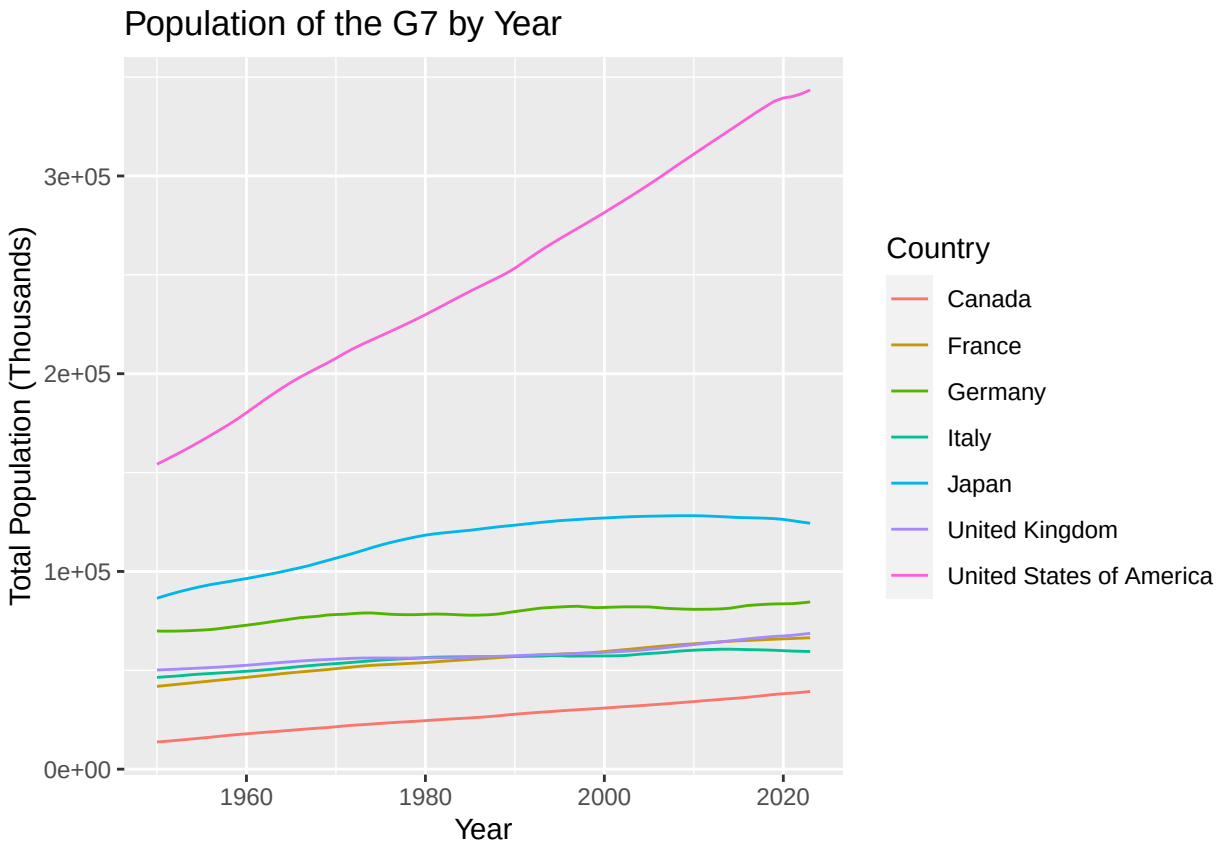
Interpretations: Diagram looks to analyze changes in population density over time for different regions. Found that many continents and the world in general will eventually peak in population and eventually start decreasing in the near future, with only continents like Oceania, North America, and Africa still increasing.

```
required_combine_data |>
  filter(type %in% c("Region", "World")) |>
  group_by(year, region_subregion_country_or_area) |>
  ggplot(aes(x = year, y = sex_ratio_at_birth_males_per_100_female_births, group = factor(region_subreg
  geom_line() +
  labs( x = "Year", y = "Sex Ratio at Birth (Males per 100 Female Births)", title = "Trends in Sex Ratio
```



Interpretations: Diagram looks to analyze how the sex ratio for certain countries changes over time. One thing to notice is that China had a very large jump in ratio in the early 2000s, which could have been due to at the time in the 1990s China seemed like they were trying to prevent population change from getting to large. However, since then, and for most countries the ratio seem to stay around the same number.

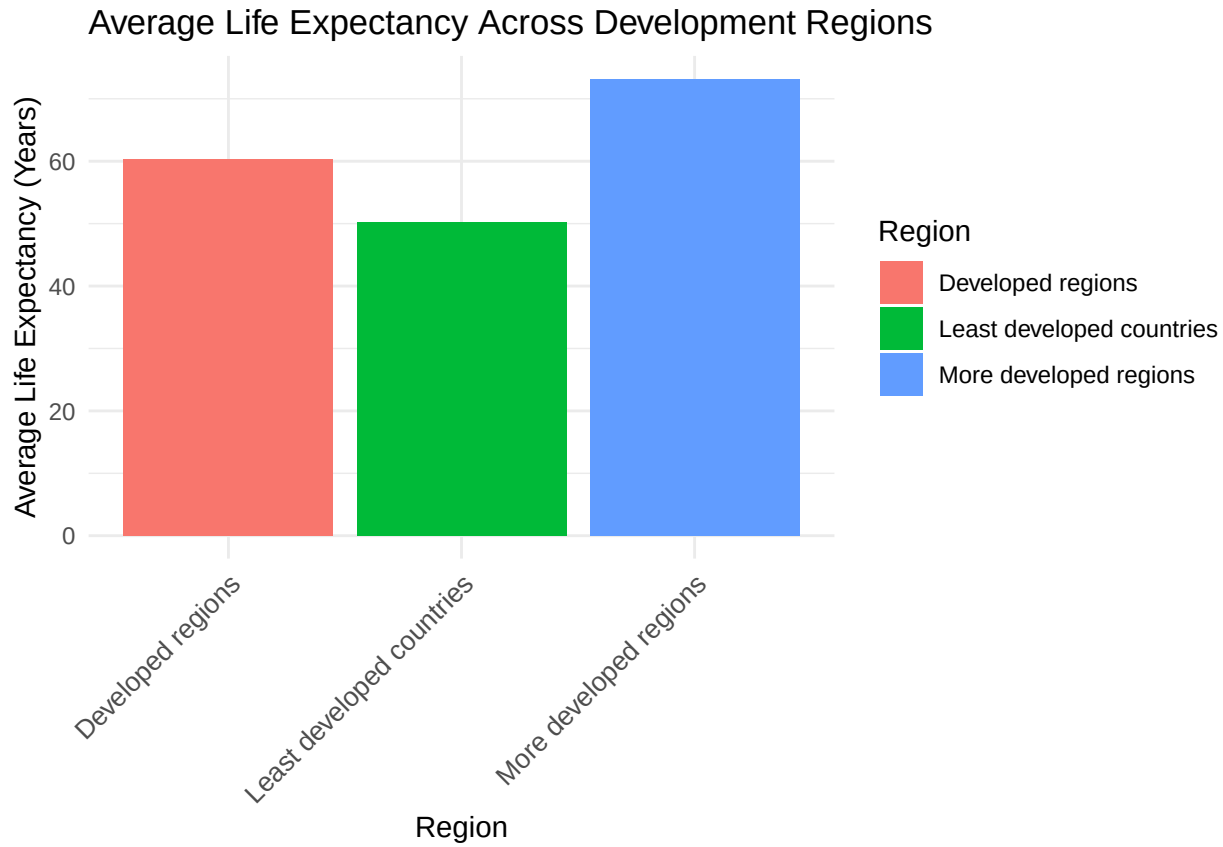
```
group_of_seven <- required_estimates_data |> filter(region_subregion_country_or_area %in% c("Canada","F

group_of_seven |>
  ggplot(aes(x = year, y = total_population_as_of_1_july_thousands, color = region_subregion_country_or
    title = "Population of the G7 by Year",
    x = "Year",
    y = "Total Population (Thousands)",
    color = "Country"
  )
```



Interpretations: Displays the population growth trends over time for each G7 country, showing how populations have changed year over year. Every country has an increase in population over time, however the United states by far has seen the largest increase in population in the group of these seven countries. Japan also has appeared to decrease in population in recent years with this seeming like it could be a common trend in the future.

```
developed_data <- required_estimates_data |>
  filter(region_subregion_country_or_area %in% c(
    "Least developed countries",
    "Less developed regions, excluding least developed countries",
    "More developed regions"
  )) |>
  mutate(
    region_subregion_country_or_area = case_when(
      region_subregion_country_or_area == "Less developed regions, excluding least developed countries"
      TRUE ~ region_subregion_country_or_area
    )
  )

developed_data_summary <- developed_data |>
  group_by(region_subregion_country_or_area) |>
  summarise(
    avg_life_expectancy = mean(life_expectancy_at_birth_both_sexes_years, na.rm = TRUE)
  )

ggplot(developed_data_summary, aes(
    y = avg_life_expectancy,
    x = region_subregion_country_or_area,
```

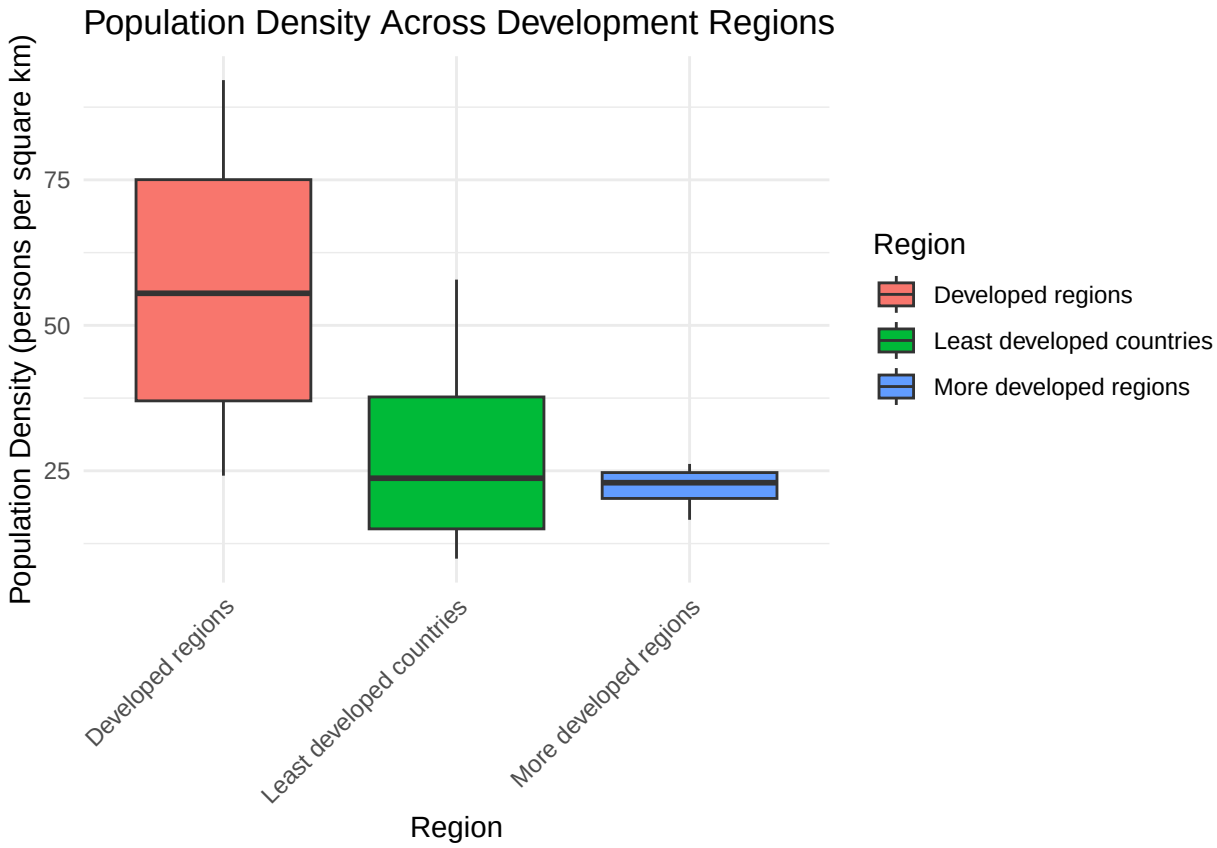
```
    fill = region_subregion_country_or_area)) +
  geom_bar(stat = "identity") +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1, size = 10)
  ) +
  labs(
    title = "Average Life Expectancy Across Development Regions",
    x = "Region",
    y = "Average Life Expectancy (Years)",
    fill = "Region"
  )
```



Interpretations: Visualizes the breakdown of life expectancy among countries that are considered to be developed. What we can see is a trend that makes sense in which the more developed a region is, the longer the life expectancy is, especially finding that there almost seems to be a similar jump in life expectancy as the development category increases.

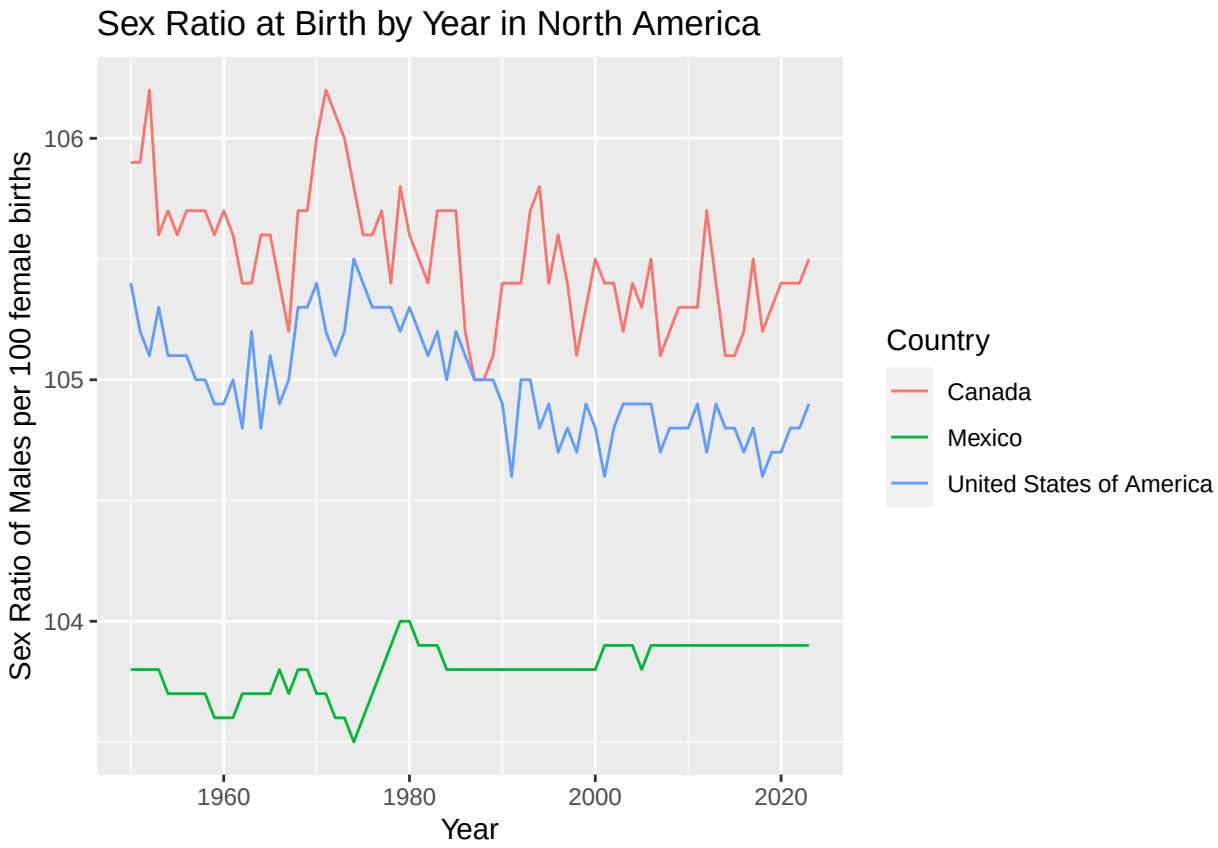
```
ggplot(developed_data, aes(
    x = region_subregion_country_or_area,
    y = population_density_as_of_1_july_persons_per_square_km,
    fill = region_subregion_country_or_area)) +
  geom_boxplot() +
  theme_minimal() +
  labs(
    title = "Population Density Across Development Regions",
    x = "Region",
    y = "Population Density (persons per square km)",
```

```
    fill = "Region"
  ) +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1)
  )
```



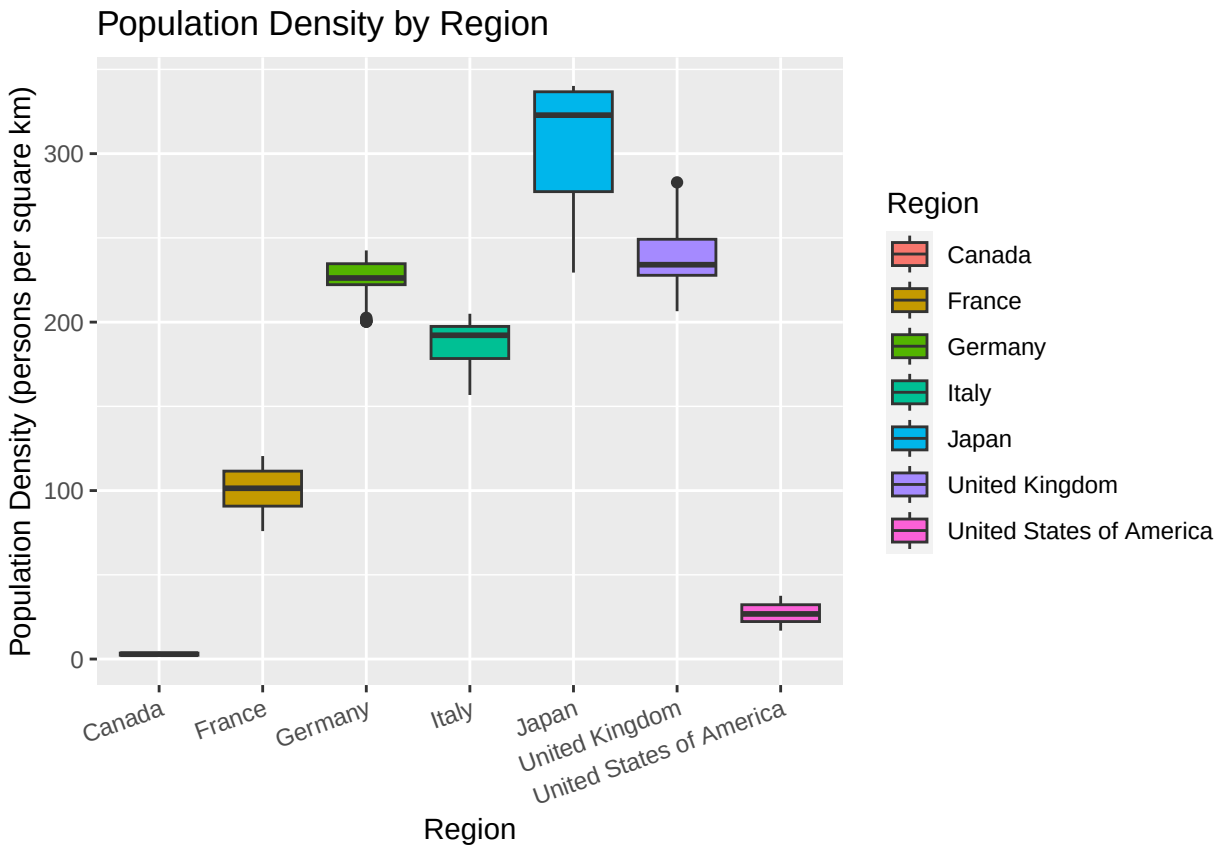
Interpretations: Provides summary statistics for the population density of countries with different types of level of development. What we can see it that while the median of population density for the developed regions is much larger than that of the other levels of development, the development regions have a very large range indicating there can be developed countries with high and low population densities. However, specifically with more developed regions, they have very little spread at all indiciating that these types of counties are almost guaranteed to have small population densities.

```
north_america <- required_estimates_data |> filter(region_subregion_country_or_area %in% c("United State

north_america |> ggplot(aes(x = year, y = sex_ratio_at_birth_males_per_100_female_births, color = regio
```



Interpretations: Visualizes the breakdown of populations by gender (Male vs. Female) for North American countries in 2023, based on the sex ratio at birth. What we can see is that Canada has a much larger percentage of males being born, and on the other side Mexico has a very little percentage of males being born in comparison to the other countries, with the USA being more of a middle ground.

```
group_of_seven |>
  ggplot(aes(x = region_subregion_country_or_area, y = population_density_as_of_1_july_persons_per_squa
  geom_boxplot() +
  labs(
    title = "Population Density by Region",
    x = "Region",
    y = "Population Density (persons per square km)",
    fill = "Region"
  ) + theme(
    axis.text.x = element_text(angle = 20, hjust = 1) )
```

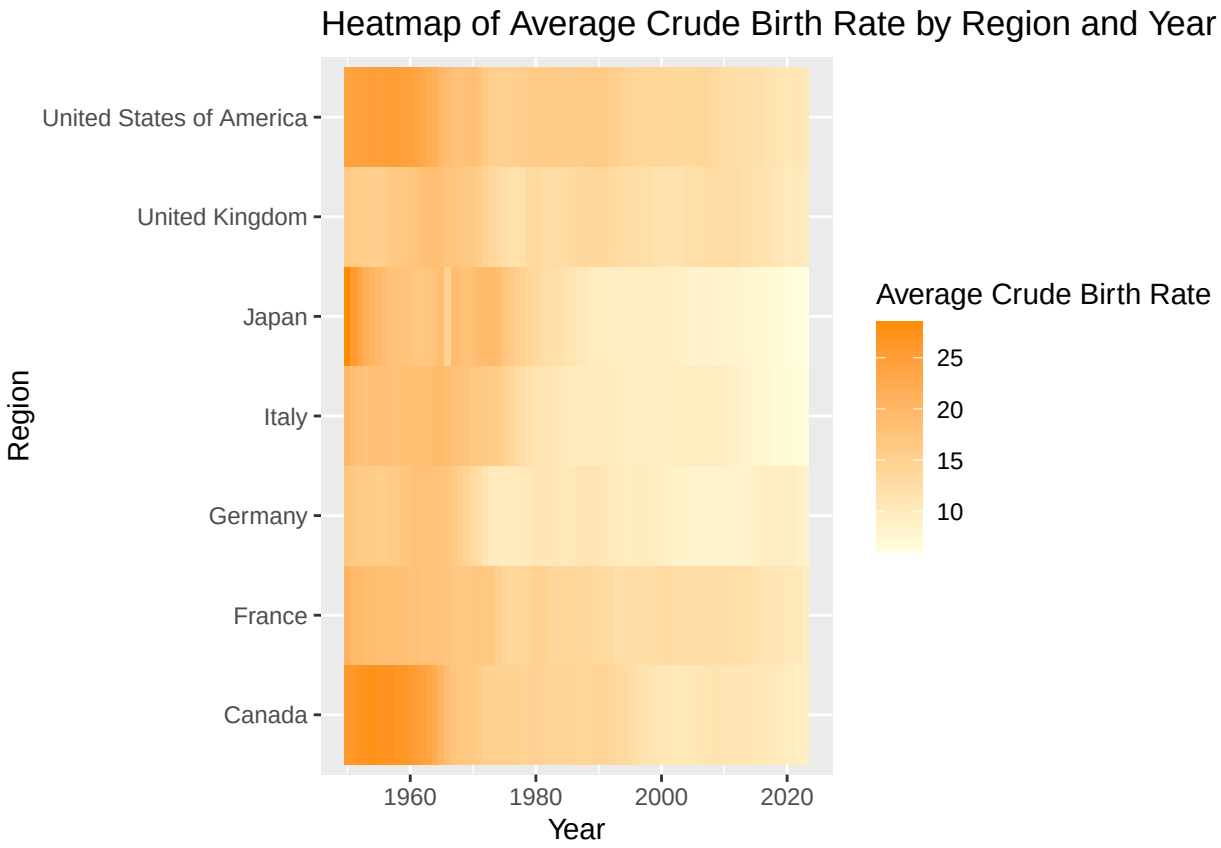


Interpretations: Shows the distribution of population densities (persons per square km) across regions, including the spread and potential outliers. Something interesting to note in this data is that Japan is incredibly dense, which could play into the drop in population projections that we saw in an earlier diagram. Also the United States of America and Canada both have very low densities in comparison to other countries which might show why they have the largest increase in population projections

```
g7_aggregated <- group_of_seven |>
  group_by(region_subregion_country_or_area, year) |>
  summarise(
    avg_crude_birth_rate = mean(crude_birth_rate_births_per_1_000_population, na.rm = TRUE)
  )
```

```
## `summarise()` has grouped output by 'region_subregion_country_or_area'. You can
## override using the `.groups` argument.
```

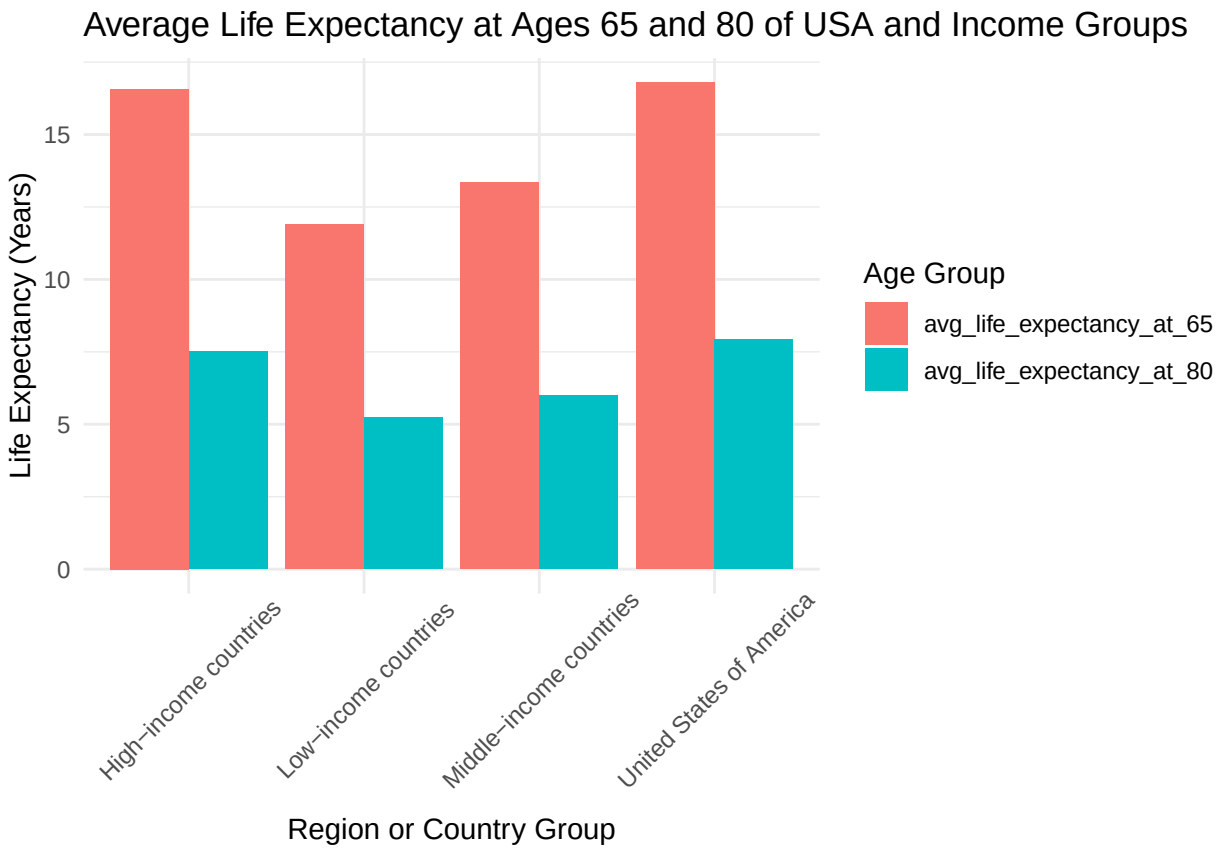
```
# Heatmap of Average Crude Birth Rate by Region and Year
g7_aggregated |>
  ggplot(aes(x = year, y = region_subregion_country_or_area, fill = avg_crude_birth_rate)) +
  geom_tile() +
  scale_fill_gradient(low = "lightyellow", high = "darkorange") +
  labs(
    title = "Heatmap of Average Crude Birth Rate by Region and Year",
    x = "Year",
    y = "Region",
    fill = "Average Crude Birth Rate"
  )
```



Interpretations: Displays a heat map illustrating the average crude birth rate over time across different regions, highlighting patterns and trends in birth rates over the years. Something that is made very clear in the charts is that every region is decreasing in birth rate, with United States and Canada seeing drastic decreases in birth rate after a boom in the 1960s.

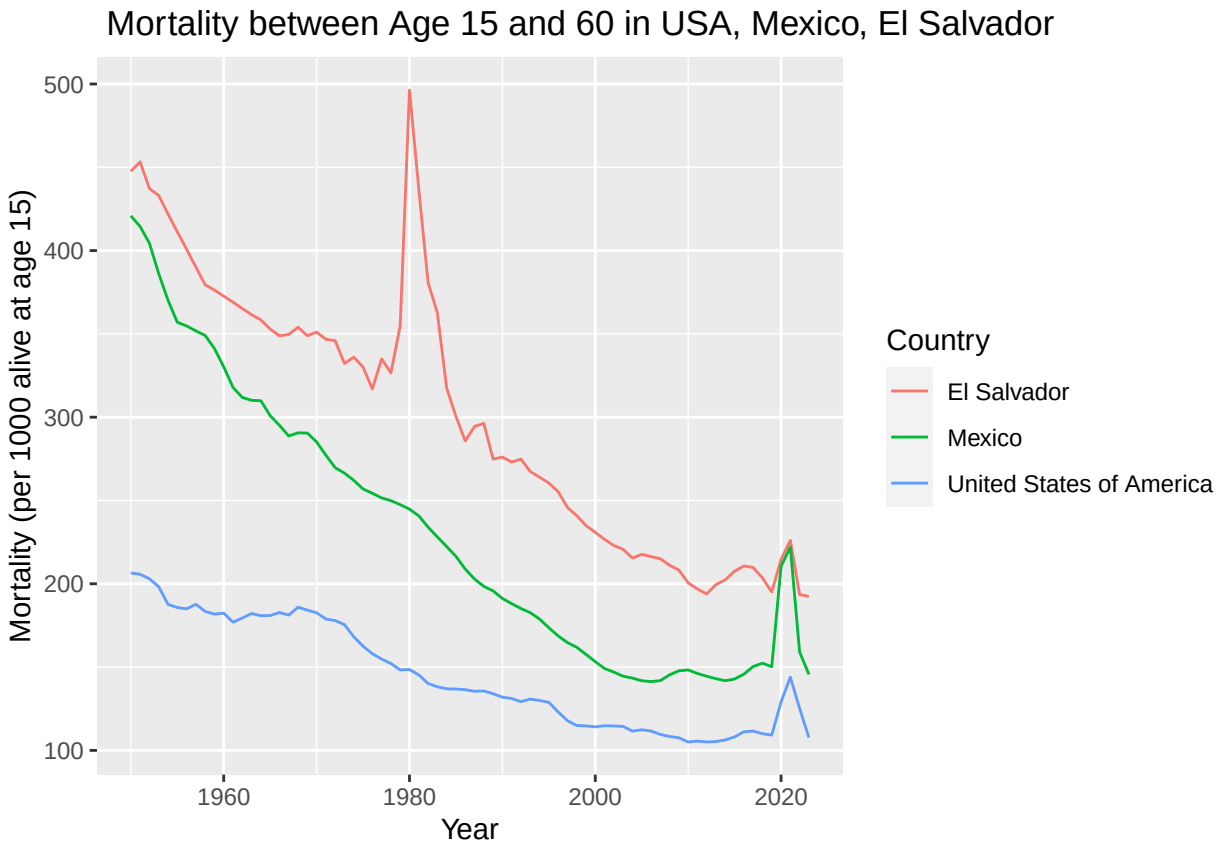
```
continent_life_expectancy <- required_estimates_data |> filter(region_subregion_country_or_area %in% c(

avg_life_expectancy <- continent_life_expectancy |>
  group_by(region_subregion_country_or_area) |>
  summarise(
    avg_life_expectancy_at_65 = mean(life_expectancy_at_age_65_both_sexes_years, na.rm = TRUE),
    avg_life_expectancy_at_80 = mean(life_expectancy_at_age_80_both_sexes_years, na.rm = TRUE)
  )

# Reshape the data to long format
life_expectancy_long <- avg_life_expectancy |>
  pivot_longer(cols = c(avg_life_expectancy_at_65, avg_life_expectancy_at_80),
               names_to = "age_group",
               values_to = "life_expectancy")

ggplot(life_expectancy_long, aes(x = region_subregion_country_or_area,
                                 y = life_expectancy,
                                 fill = age_group)) +
  geom_bar(stat = "identity", position = position_dodge()) +
  labs(
    title = "Average Life Expectancy at Ages 65 and 80 of USA and Income Groups",
    x = "Region or Country Group",
    y = "Life Expectancy (Years)",
```

```
    fill = "Age Group"
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, vjust = 0.6)  # Rotate x-axis labels vertically
  )
```



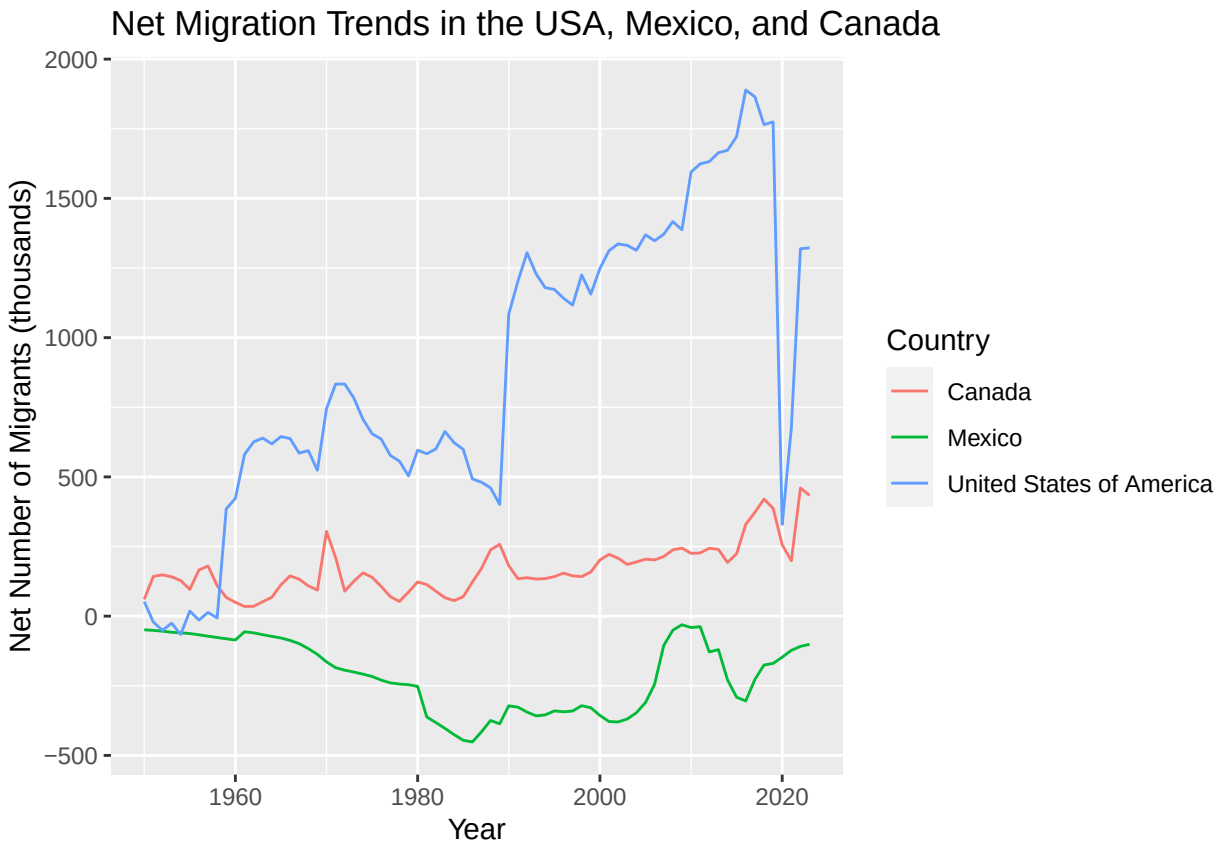
Interpretations: Compares life expectancy at older ages between different incomes of countries. Countries that have high-income natrually have a higher life expectancy for both 65 year olds and 80 year olds, but see a very massive jump for 65 year olds indicating that 65 tends to be a time where people begin reaching death in low or middle age countries while in high income countries like the United States they still have some strong years left.

```
usa_salvador_mex <- required_estimates_data |> filter(region_subregion_country_or_area %in% c("United S

usa_salvador_mex |> ggplot(aes(x=year,
                           y=mortality_between_age_15_and_60_both_sexes_deaths_under_age_60_per_1_000_al
                           color=region_subregion_country_or_area)) + geom_line() + labs(title = " Morta
```



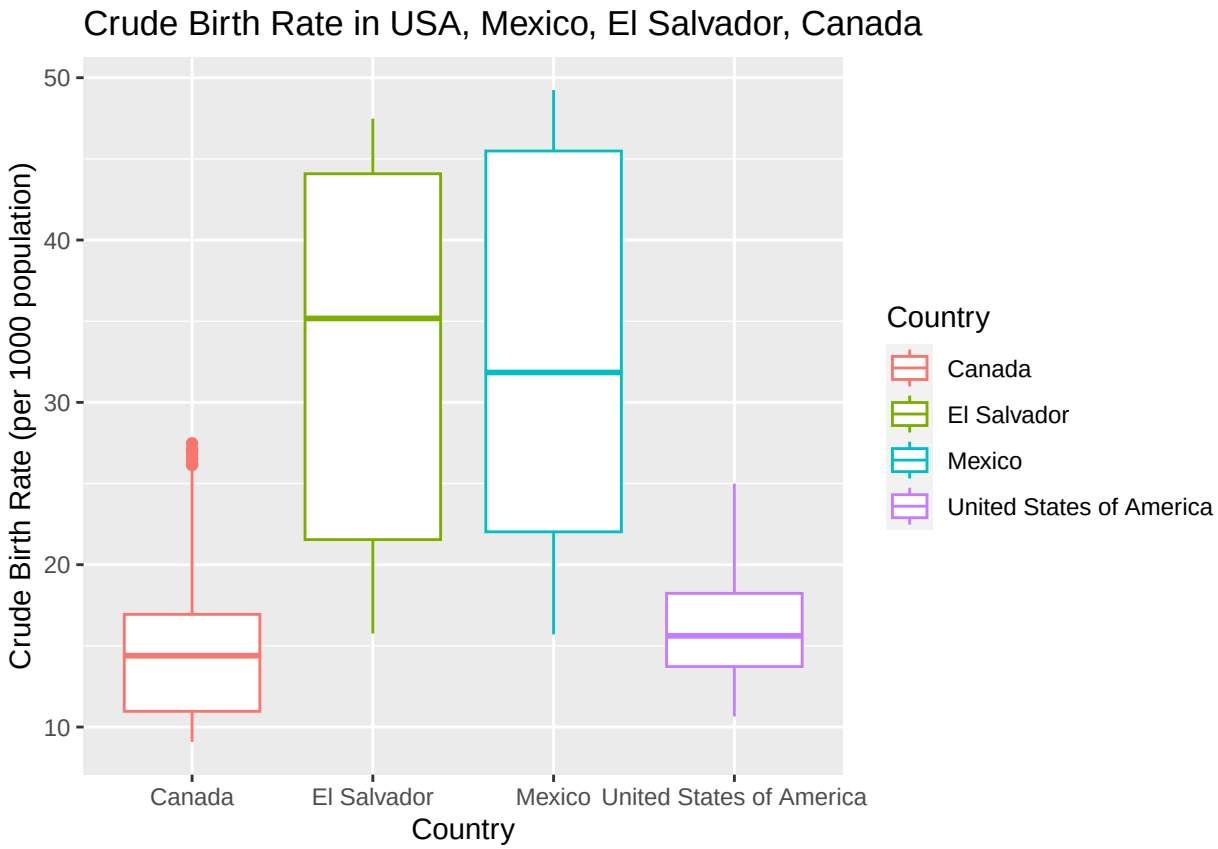
Interpretations: Compares the mortality rate among North America, United States, and El Salvador. As we can see overall every country has dropped in mortality rate, however there have been large spikes such as the Covid pandemic which causes the spikes in 2020 and the civil war in El Salvador which shows a massie spike the the 1980s

```
migration <- required_estimates_data |> filter(region_subregion_country_or_area %in% c("United States o

migration |> ggplot(aes(x = year,
                        y = net_number_of_migrants_thousands,
                        color = region_subregion_country_or_area)) +
  geom_line() +
  labs(
    title = "Net Migration Trends in the USA, Mexico, and Canada",
    x = "Year",
    y = "Net Number of Migrants (thousands)",
    color = "Country"
  )
```



Interpretations: Looks into detail at the difference in number of migrants among North American countries. Something very interesting to note here is that the United States has had individual large jumps in it's number of migrants over the years, with Canada and the Mexico staying fairly constant over time. One thing that was surprising to see in this diagram is that Mexico didn't appear to drop in migrants in 2020 but rather increase which is weird to see considering the conditions of the pandemic.

```
usa_salvador_mex <- required_estimates_data |> filter(region_subregion_country_or_area %in% c("United S

usa_salvador_mex |> group_by(region_subregion_country_or_area) |>
  ggplot(aes(x=region_subregion_country_or_area,
             y=crude_birth_rate_births_per_1_000_population,
             color=region_subregion_country_or_area)) + geom_boxplot() + labs(title = "Crude Birth Rate
```



Interpretations: A diagram to compare summary statistics for the birth rates in different North American countries as well as El Salvador. What we can interpret is that the birth rates in the United States and Canada stay fairly low with little variability, and even though the median birth rate in both El Salvador and Mexico are high in comparison they have large variability which indicates that the birth rate is constantly changing from year to year.

5. Requirement-5 (2 pt) Having developed a strong understanding of your data, you'll now create a machine learning (ML) model to predict a specific metric. This involves selecting the most relevant variables from your dataset.

The UN's World Population Prospects provides a range of projected scenarios of population change. These rely on different assumptions in fertility, mortality and/or migration patterns to explore different demographic futures. Check this link for more info: https://population.un.org/wpp/DefinitionOfProjectionScenarios

You can choose to predict the same metric the UN provides (e.g., future population using fertility, mortality, and migration data). Compare your model's predictions to the UN's.

How significantly do your population projections diverge from those of the United Nations? Provide a comparison of the two. If you choose a different projection for which there is no UN data to compare with, then this comparison is not required.

```
library(tidyverse)
library(caret)
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:purrr':
##
```

```
##     lift
```

```
# Step 1: Simplify and Clean Data
training_data <- required_estimates_data %>%
  filter(region_subregion_country_or_area == "World") %>%
  select(
    year,
    total_population_as_of_1_july_thousands,
    total_fertility_rate_live_births_per_woman,
    life_expectancy_at_birth_both_sexes_years,
    net_number_of_migrants_thousands
  ) %>%
  drop_na()

# Bind the data for training and testing
train_data <- training_data %>% filter(year <= 2010)
test_data <- training_data %>% filter(year > 2010)

# Step 2: Build the Linear Regression Model
population_model <- lm(
  total_population_as_of_1_july_thousands ~ total_fertility_rate_live_births_per_woman +
    life_expectancy_at_birth_both_sexes_years +
    net_number_of_migrants_thousands,
  data = train_data
)

# Model Summary
cat("Model Summary:\n")
```

```
## Model Summary:
```

```
summary(population_model)
```

```
##
## Call:
## lm(formula = total_population_as_of_1_july_thousands ~ total_fertility_rate_live_births_per_woman +
##     life_expectancy_at_birth_both_sexes_years + net_number_of_migrants_thousands,
##     data = train_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -328700 -148501   22854  126037  400770
##
## Coefficients: (1 not defined because of singularities)
##                                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                               -373052     956477  -0.390    0.698
## total_fertility_rate_live_births_per_woman  -606504      80346  -7.549 3.51e-10
## life_expectancy_at_birth_both_sexes_years    122304      11073  11.045 6.86e-16
## net_number_of_migrants_thousands                 NA         NA      NA       NA
##
## (Intercept)
## total_fertility_rate_live_births_per_woman ***
## life_expectancy_at_birth_both_sexes_years  ***
## net_number_of_migrants_thousands
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 193900 on 58 degrees of freedom
## Multiple R-squared:  0.981,  Adjusted R-squared:  0.9804
## F-statistic:  1499 on 2 and 58 DF,  p-value: < 2.2e-16
```

```
# Evaluate Model Performance on Training Data
cat("R-squared (Training Data):", summary(population_model)$r.squared, "\n")
```

```
## R-squared (Training Data): 0.9810263
```

```
cat("Adjusted R-squared (Training Data):", summary(population_model)$adj.r.squared, "\n")
```

```
## Adjusted R-squared (Training Data): 0.9803721
```

```
# Step 3: Predict on Test Data and Evaluate
test_data <- test_data %>%
  mutate(predicted_population = predict(population_model, newdata = .))
```

```
## Warning: There was 1 warning in `mutate()`.
## i In argument: `predicted_population = predict(population_model, newdata = .)`.
## Caused by warning in `predict.lm()`:
## ! prediction from rank-deficient fit; attr(*, "non-estim") has doubtful cases
```

```
# Calculate R-squared and MSE using caret package
model_performance <- postResample(pred = test_data$predicted_population, obs = test_data$total_populati
cat("R-squared (Test Data):", model_performance["Rsquared"], "\n")
```

```
## R-squared (Test Data): 0.876298
```

```
cat("Mean Squared Error (Test Data):", model_performance["RMSE"]^2, "\n")
```

```
## Mean Squared Error (Test Data): 534364093754
```

```
# Residuals Analysis (optional but useful for insight)
residuals <- test_data$total_population_as_of_1_july_thousands - test_data$predicted_population
cat("Residuals Summary:\n")
```

```
## Residuals Summary:
```

```
summary(residuals)
```

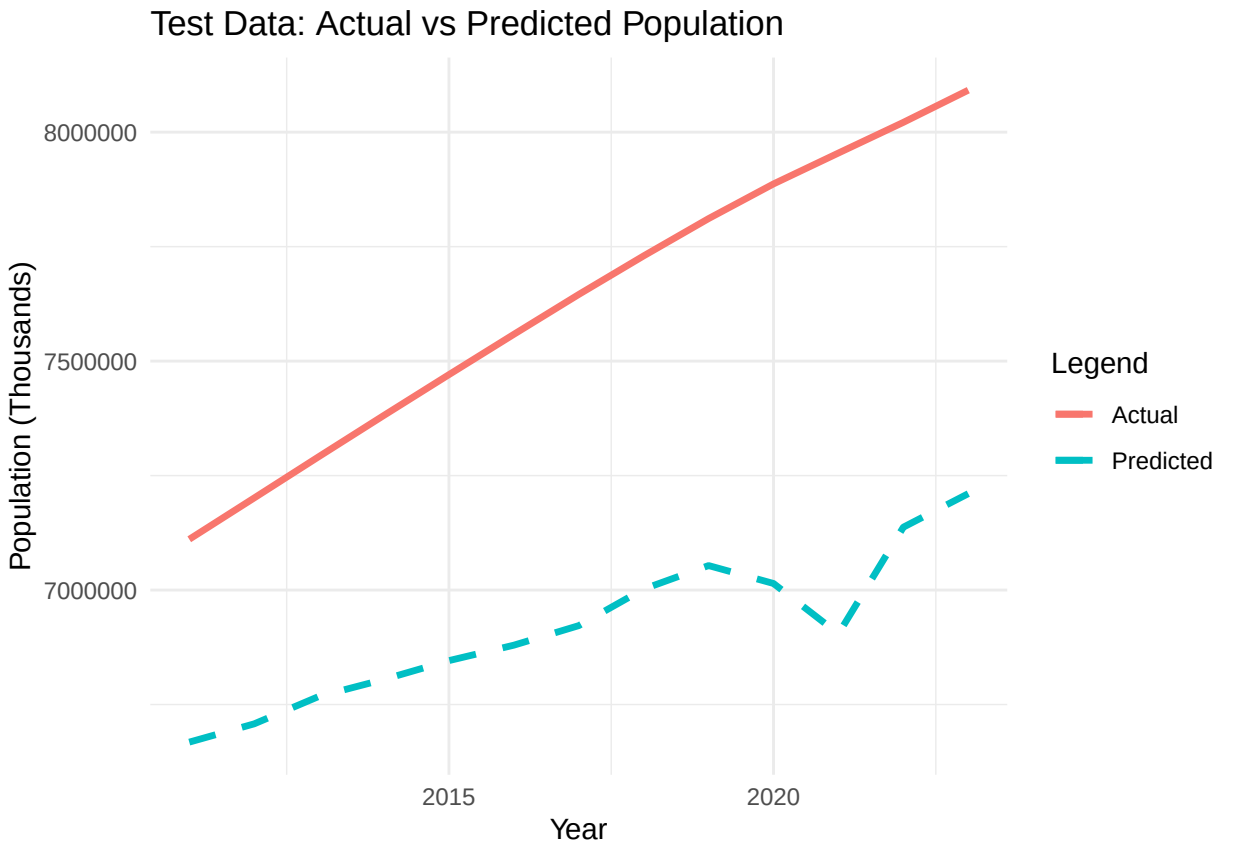
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  443067  576778  723302  710552  872647 1049956
```

```
# Step 4: Compare Actual vs Predicted on Test Data
comparison_test <- test_data %>%
  ggplot(aes(x = year)) +
  geom_line(aes(y = total_population_as_of_1_july_thousands, color = "Actual"), size = 1.2) +
  geom_line(aes(y = predicted_population, color = "Predicted"), size = 1.2, linetype = "dashed") +
  labs(
    title = "Test Data: Actual vs Predicted Population",
    x = "Year",
    y = "Population (Thousands)",
    color = "Legend"
  ) +
  theme_minimal()
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
```

```
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
print(comparison_test)
```




```
# Step 5: Predict Future Values (2024-2100)
future_data <- tibble(
  year = 2024:2100,
  total_fertility_rate_live_births_per_woman = seq(2.5, 1.5, length.out = 77),  # Example assumption
  life_expectancy_at_birth_both_sexes_years = seq(70, 85, length.out = 77),
  net_number_of_migrants_thousands = seq(1e3, 5e3, length.out = 77)  # Example assumption
)

# Predict future populations
future_predictions <- predict(population_model, newdata = future_data)
```

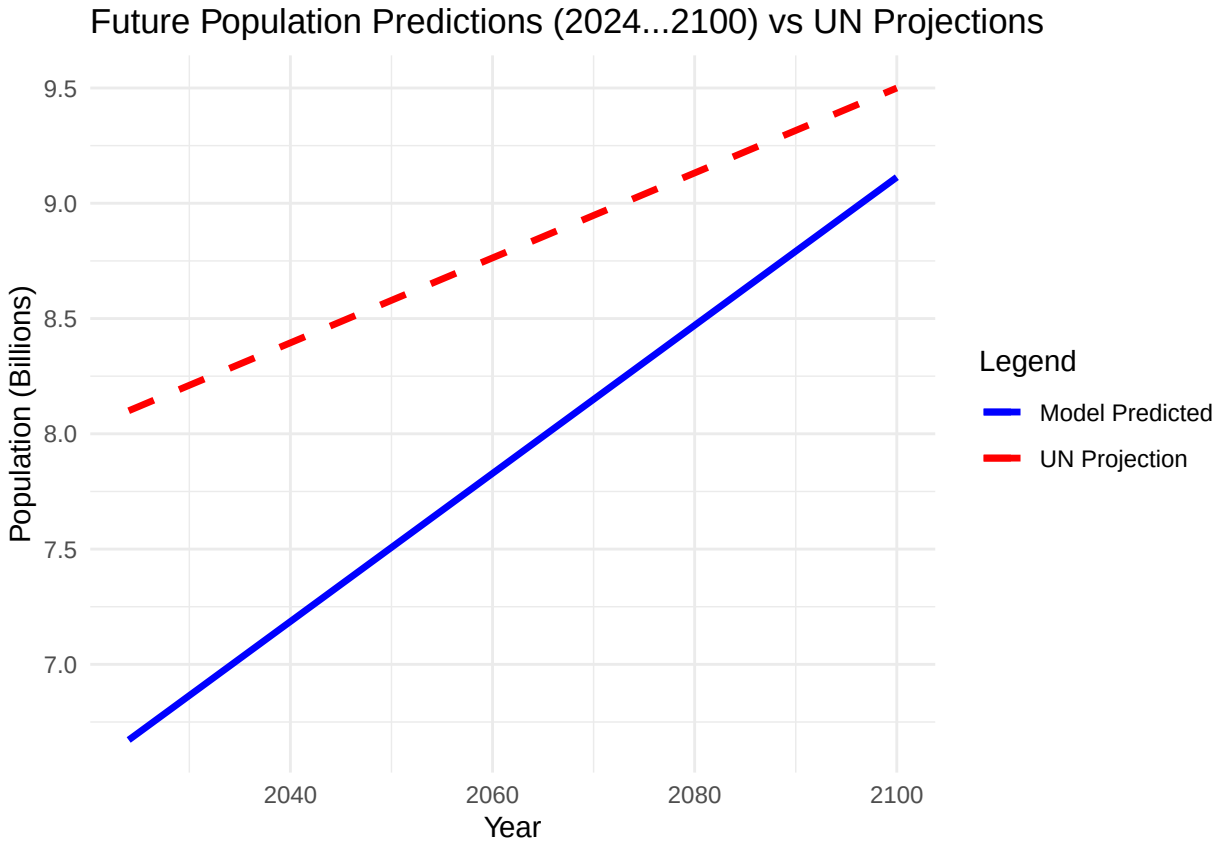
```
## Warning in predict.lm(population_model, newdata = future_data): prediction from
## rank-deficient fit; attr(*, "non-estim") has doubtful cases
```

```
# Assume UN population projections (example values, replace with actual UN projections)
un_projection_data <- tibble(
  year = 2024:2100,
  un_predicted_population = seq(8100000, 9500000, length.out = 77)
)

# Combine the predicted populations from the model and the UN projections
combined_predictions <- left_join(future_data, un_projection_data, by = "year") %>%
  mutate(predicted_population = future_predictions)
```

```r
# Visualize both Model Predictions and UN Projections
future_plot <- ggplot(combined_predictions, aes(x = year)) +
  geom_line(aes(y = predicted_population / 1e6, color = "Model Predicted"), size = 1.2) +
  geom_line(aes(y = un_predicted_population / 1e6, color = "UN Projection"), size = 1.2, linetype = "da
  labs(
    title = "Future Population Predictions (2024-2100) vs UN Projections",
    x = "Year",
    y = "Population (Billions)",
    color = "Legend"
  ) +
  theme_minimal() +
  scale_color_manual(values = c("Model Predicted" = "blue", "UN Projection" = "red"))

print(future_plot)
```

```
## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Future Population Predictions (2024-2100) vs UN
## Projections' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Future Population Predictions (2024-2100) vs UN
## Projections' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Future Population Predictions (2024-2100) vs UN
## Projections' in 'mbcsToSbcs': dot substituted for <93>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Future Population Predictions (2024-2100) vs UN
## Projections' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Future Population Predictions (2024-2100) vs UN
## Projections' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Future Population Predictions (2024-2100) vs UN
## Projections' in 'mbcsToSbcs': dot substituted for <93>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Future Population Predictions (2024-2100) vs UN
## Projections' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Future Population Predictions (2024-2100) vs UN
## Projections' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Future Population Predictions (2024-2100) vs UN
## Projections' in 'mbcsToSbcs': dot substituted for <93>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Future Population Predictions (2024-2100) vs UN
## Projections' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Future Population Predictions (2024-2100) vs UN
## Projections' in 'mbcsToSbcs': dot substituted for <80>
```

```
## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Future Population Predictions (2024-2100) vs UN
## Projections' in 'mbcsToSbcs': dot substituted for <93>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Future Population Predictions (2024-2100) vs UN
## Projections' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Future Population Predictions (2024-2100) vs UN
## Projections' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Future Population Predictions (2024-2100) vs UN
## Projections' in 'mbcsToSbcs': dot substituted for <93>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Future Population Predictions (2024-2100) vs UN
## Projections' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Future Population Predictions (2024-2100) vs UN
## Projections' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Future Population Predictions (2024-2100) vs UN
## Projections' in 'mbcsToSbcs': dot substituted for <93>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Future Population Predictions (2024-2100) vs UN
## Projections' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Future Population Predictions (2024-2100) vs UN
## Projections' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Future Population Predictions (2024-2100) vs UN
## Projections' in 'mbcsToSbcs': dot substituted for <93>

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Future Population Predictions (2024-2100) vs UN
## Projections' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Future Population Predictions (2024-2100) vs UN
## Projections' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Future Population Predictions (2024-2100) vs UN
## Projections' in 'mbcsToSbcs': dot substituted for <93>
```



6. Requirement-6 (1 pt)

Conclusion

Through our analytic work, we were able to learn a lot about the populations in countries and regions across the world. The first thing that we focused on finding in our data was to look through the population trends and find in what regions was it more common to see population trending downwards or upwards. We found that the continents of Asia, Europe, and South America have projections for a plateau and decrease, while on the other hand North America, Oceania, and Africa are still projected to increase. However, with the entire world projected to eventually plateau and decrease in population, it appears that the countries with projected drops in population are going to be more severe than those with projected increases. Another piece of information that we were able to learn through our analysis was that Japan will have one of the more severe drops in population in the future especially among the G7 countries. A possible explanation for this drop in population in Japan is likely due to the densities of the countries, as we also found through our work that Japan's population density is much larger than that of other countries, especially countries like Canada and the United States who have low population densities and are projected to still increase in population. This would indicate that because Japan is running so low on room, that the population will likely have to decrease in the future because there won't be room left. Something else that is interesting to note regarding population changes is that high-income countries like Canada and the United States have very high life expectancy, which might partly explain why those countries have projections of population increase because these countries won't be seeing the same drops in population in the elderly that other type of countries will. Finally, one more insight we were able to uncover through our work is regarding the effect of the Covid 19 pandemic on countries populations. One that was interesting to see is that naturally the pandemic caused decreases in the number of migrations for most countries in North America, however Mexico not only didn't see a decrease but rather saw a slight increase which is weird to see in a time where people naturally stopped migrating because of the issues with every country around the world.

7. Extra Credit (1 pt) Develop an interactive Shiny app to visualize your machine learning model's projections. The app must include at least one interactive widget (e.g., dropdown, radio buttons, text input)

allowing users to select a variable value (such as country/region) and view the corresponding projections.