DATASCI 306, Fall 2024, Final Group Project

Shalini Asokkumar, Anusha Chinthamaduka, Sophia Giuliani, Jonathan Sarasa, Alicia Zhou

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_excel("data.xlsx", skip = 15, sheet = "Estimates")
```

```
## New names:
## * `` -> `...1`
## * `` -> `...2`
## * `` -> `...3`
## * `` -> `...4`
## * `` -> `...5`
## * `` -> `...6`
## * `` -> `...7`
## * `` -> `...8`
## * `` -> `...9`
## * `` -> `...10`
## * `` -> `...11`
## * `` -> `...13`
## * `` -> `...14`
## * `` -> `...15`
## * `` -> `...16`
```

```
## * `` -> `...17`
## * `` -> `...18`
## * `` -> `...19`
## * `` -> `...20`
## * `` -> `...21`
## * `` -> `...22`
## * `` -> `...23`
## * `` -> `...25`
## * `` -> `...26`
## * `` -> `...27`
## * `` -> `...28`
## * `` -> `...29`
## * `` -> `...30`
## * `` -> `...32`
## * `` -> `...33`
## * `` -> `...34`
## * `` -> `...35`
## * `` -> `...36`
## * `` -> `...37`
## * `` -> `...38`
## * `` -> `...39`
## * `` -> `...40`
## * `` -> `...41`
## * `` -> `...42`
## * `` -> `...43`
## * `` -> `...44`
## * `` -> `...45`
## * `` -> `...46`
## * `` -> `...47`
## * `` -> `...48`
## * `` -> `...49`
## * `` -> `...50`
## * `` -> `...51`
## * `` -> `...52`
## * `` -> `...53`
## * `` -> `...54`
## * `` -> `...55`
## * `` -> `...56`
## * `` -> `...57`
## * `` -> `...58`
## * `` -> `...59`
## * `` -> `...60`
## * `` -> `...61`
## * `` -> `...62`
## * `` -> `...63`
## * `` -> `...65`
```

```r
colnames(estimates) <- as.character(unlist(estimates[1, ]))
estimates = estimates[-1,]

mediums = read_excel("data.xlsx", skip = 15, sheet = "Medium variant")
```

```
## New names:
## * `` -> `...1`
## * `` -> `...2`
```

```
## * `` -> `...3`
## * `` -> `...4`
## * `` -> `...5`
## * `` -> `...6`
## * `` -> `...7`
## * `` -> `...8`
## * `` -> `...9`
## * `` -> `...10`
## * `` -> `...11`
## * `` -> `...13`
## * `` -> `...14`
## * `` -> `...15`
## * `` -> `...16`
## * `` -> `...17`
## * `` -> `...18`
## * `` -> `...19`
## * `` -> `...20`
## * `` -> `...21`
## * `` -> `...22`
## * `` -> `...23`
## * `` -> `...25`
## * `` -> `...26`
## * `` -> `...27`
## * `` -> `...28`
## * `` -> `...29`
## * `` -> `...30`
## * `` -> `...32`
## * `` -> `...33`
## * `` -> `...34`
## * `` -> `...35`
## * `` -> `...36`
## * `` -> `...37`
## * `` -> `...38`
## * `` -> `...39`
## * `` -> `...40`
## * `` -> `...41`
## * `` -> `...42`
## * `` -> `...43`
## * `` -> `...44`
## * `` -> `...45`
## * `` -> `...46`
## * `` -> `...47`
## * `` -> `...48`
## * `` -> `...49`
## * `` -> `...50`
## * `` -> `...51`
## * `` -> `...52`
## * `` -> `...53`
## * `` -> `...54`
## * `` -> `...55`
## * `` -> `...56`
## * `` -> `...57`
## * `` -> `...58`
## * `` -> `...59`
```

```
## * `` -> `...60`
## * `` -> `...61`
## * `` -> `...62`
## * `` -> `...63`
## * `` -> `...65`
```

```
colnames(mediums) <- as.character(unlist(mediums[1, ]))
mediums = mediums[-1,]
```

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.

```
est_values <- estimates |>
  select(Index, Year:last_col()) |>
  mutate(across(where(is.character), as.double, .names = '{col}'))
```

```
## Warning: There were 54 warnings in `mutate()`.
## The first warning was:
## i In argument: `across(where(is.character), as.double, .names = "{col}")`.
## Caused by warning:
## ! NAs introduced by coercion
## i Run `dplyr::last_dplyr_warnings()` to see the 53 remaining warnings.
```

```
  # converts all numeric columns from characters into doubles
  # removes Label/Separator type
estimates$Index <- estimates$Index |> as.double()
  # creates key column of Index and converts to double
estimates <- estimates |> select(Index, `Region, subregion, country or area *`,
                                 Type)
  # selecting relevant columns from original database
estimates <- estimates |> full_join(est_values, join_by(Index)) |> filter(Type != "Label/Separator")
  # only includes relevant data in Estimates

med_values <- mediums |>
  select(Index, Year:last_col()) |>
  mutate(across(where(is.character), as.double, .names = '{col}'))
```

```
## Warning: There were 54 warnings in `mutate()`.
## The first warning was:
## i In argument: `across(where(is.character), as.double, .names = "{col}")`.
## Caused by warning:
## ! NAs introduced by coercion
## i Run `dplyr::last_dplyr_warnings()` to see the 53 remaining warnings.
```

```r
mediums$Index <- mediums$Index |> as.double()
mediums <- mediums |> select(Index, `Region, subregion, country or area *`,
                                 Type)
mediums <- mediums |> full_join(med_values, join_by(Index)) |> filter(Type != "Label/Separator")
  # replicated the above for the mediums dataset

#renaming columns
corrected_colnames <- c(
  "index",
  "region_subregion_country_area",
  "type",
  "year",
  "total_pop_january_thousands",
  "total_pop_july_thousands",
  "male_pop_july_thousands",
  "female_pop_july_thousands",
  "pop_density_july_person_per_sq_km",
  "pop_sex_ratio_july_males_per_100_females",
  "med_age_july_years",
  "natural_change_births_minus_deaths_thousands",
  "rate_of_natural_change_per_1000",
  "population_change_thousands",
  "population_growth_rate_percentage",
  "population_annual_doubling_time_years",
  "births_thousands",
  "births_by_woman_aged_15_to_19_thousands",
  "crude_birth_rate_per_1000_pop",
  "total_fertility_rate_live_births_per_woman",
  "net_reproduction_rate_surviving_daughters_per_woman",
  "mean_age_childbearing_years",
  "sex_ratio_at_birth_males_per_100_female_births",
  "total_deaths_thousands",
  "male_deaths_thousands",
  "female_deaths_thousands",
  "crude_death_rate_deaths_per_1000_population",
  "total_life_expectancy_at_birth_years",
  "male_life_expectancy_at_birth_years",
  "female_life_expectancy_at_birth_years",
  "total_life_expectancy_at_age_15_years",
  "male_life_expectancy_at_age_15_years",
  "female_life_expectancy_at_age_15_years",
  "total_life_expectancy_at_age_65_years",
  "male_life_expectancy_at_age_65_years",
  "female_life_expectancy_at_age_65_years",
  "total_life_expectancy_at_age_80_years",
  "male_life_expectancy_at_age_80_years",
  "female_life_expectancy_at_age_80_years",
  "infant_deaths_under_age_1_thousands",
  "infant_morality_rate_infant_deaths_per_1000_births",
  "live_births_surviving_to_age_1_thousands",
  "under_five_deaths_thousands",
  "deaths_under_age_5_per_1,000_live_births",
  "total_male_mortality_before_age_40_per_1000_births",
```

```r
  "male_mortality_before_age_40_per_1000_births",
  "female_mortality_before_age_40_per_1000_births",
  "total_mortality_before_age_60_per_1000_births",
  "male_mortality_before_age_60_per_1000_births",
  "female_mortality_before_age_60_per_1000_births",
  "deaths_under_age_50_per_1000_total_alive_at_15",
  "deaths_under_age_50_per_1000_males_alive_at_15",
  "deaths_under_age_50_per_1000_females_alive_at_15",
  "deaths_under_age_60_per_1000_total_alive_at_15",
  "deaths_under_age_60_per_1000_males_alive_at_15",
  "deaths_under_age_60_per_1000_females_alive_at_15",
  "net_num_migrants_thousands",
  "net_migration_rate_per_1000"
)

colnames(estimates) <- corrected_colnames
colnames(mediums) <- corrected_colnames

# selecting only columns that we use in our replication and EDA

estimates <- estimates |> select(index, year, region_subregion_country_area, type, total_pop_january_tho

mediums <- mediums |> select(index, year, region_subregion_country_area, type, total_pop_january_thousa
```

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.

- population projections from 2024
- projections broken down by world and continent
- fertility rate in children/woman from 1950 - 2100
- population 1950 to 2100
- life expectancy from 1950 to 2023
- annual net migration 1950 to 2023

1. **Diagram 1:** Population Projections from 2024 - Sophia Giuliani

```r
mediums %>% select(year) %>% filter(!is.na(year)) %>% range() # 2024 - 2100
```

```
## [1] 2024 2100
```

```r
estimates %>% select(year) %>% filter(!is.na(year)) %>% range() # 1950 - 2023
```

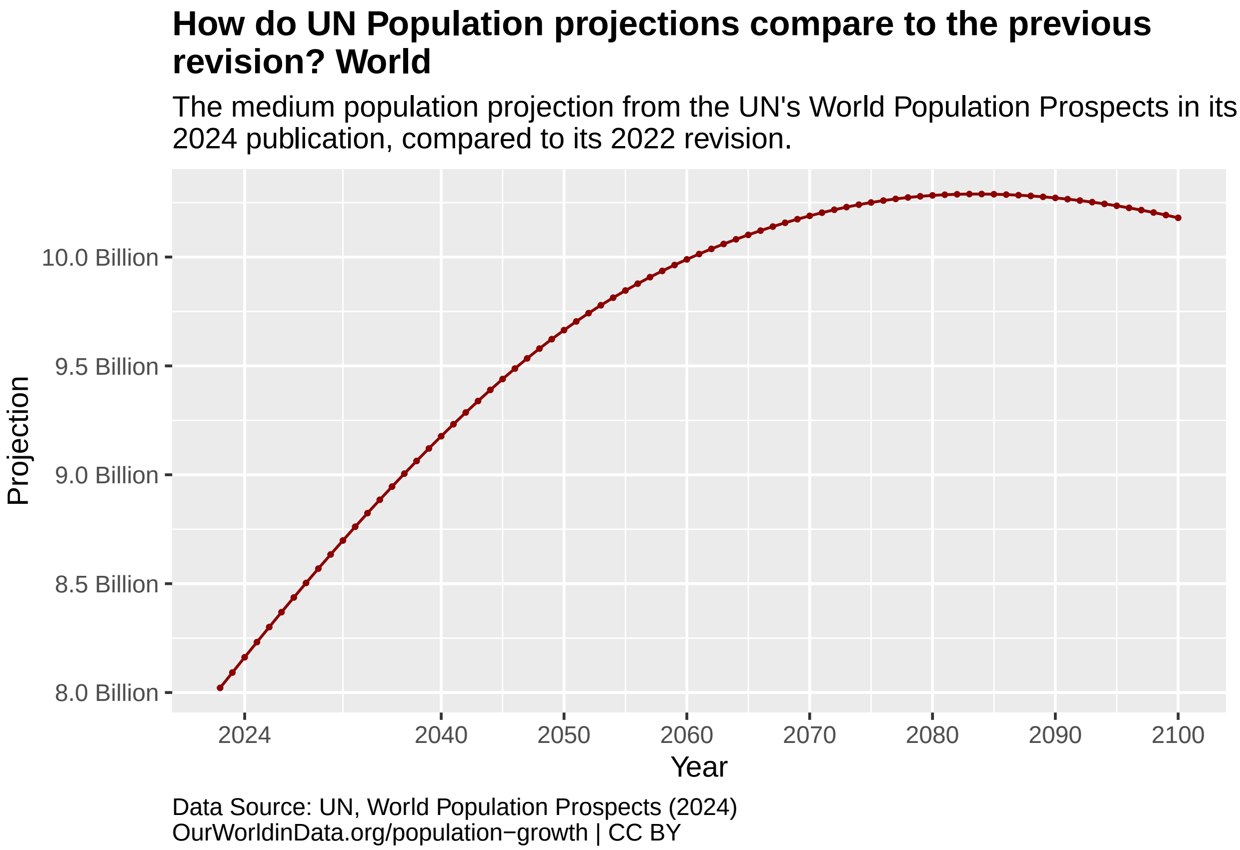
```
## [1] 1950 2023
```

```r
# Need to bind the rows from the two dataframes
combined <- rbind(estimates, mediums)
combined %>% select(year) %>% filter(!is.na(year)) %>% range() # 1950 - 2100
```

```
## [1] 1950 2100
```

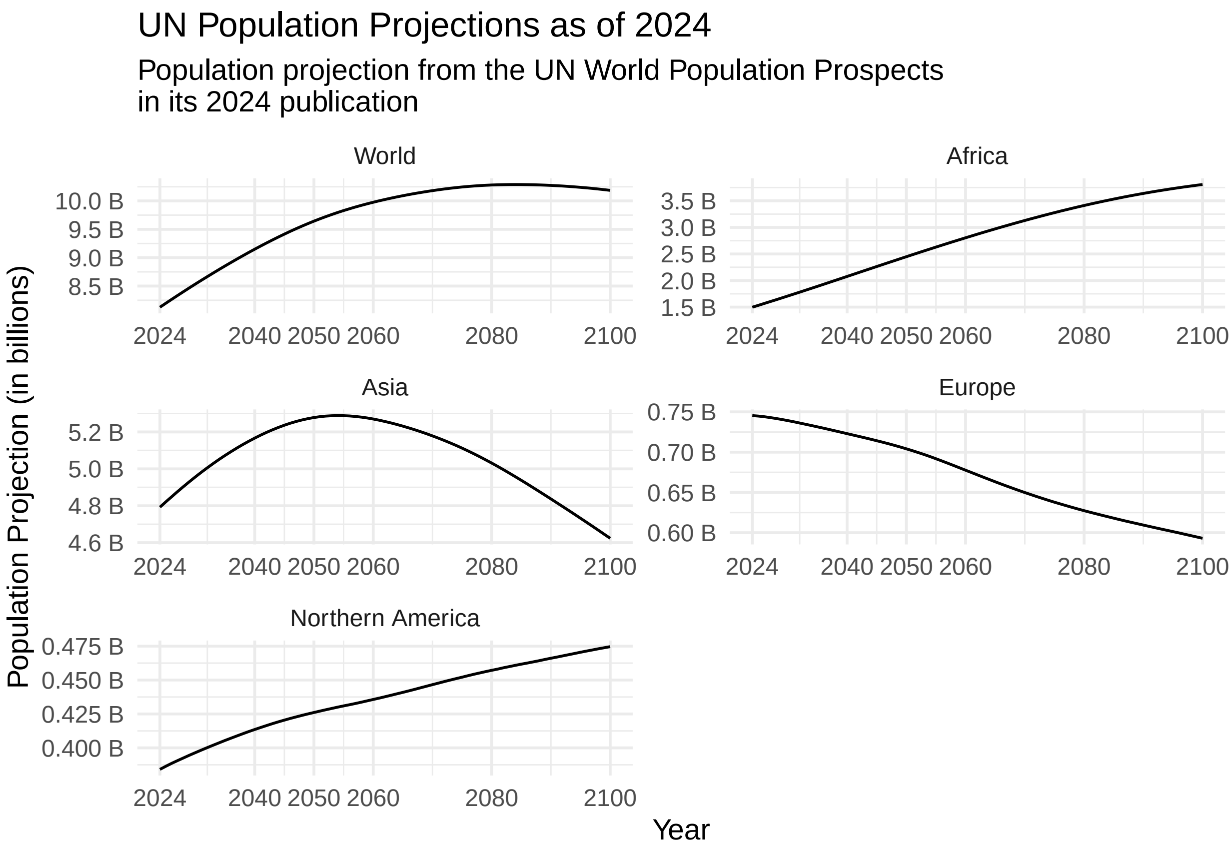
```r
combined %>%
  select(year,
         region_subregion_country_area,
         total_pop_july_thousands) %>%
  filter(!is.na(year)) %>%
  filter(year >= "2022") %>%
```

```r
filter(region_subregion_country_area == "World") %>%
ggplot(mapping = aes(x = year,
                     y = total_pop_july_thousands)) +
geom_line(color = "darkred") +
geom_point(color = "darkred",
           size = 0.5) +
scale_x_continuous(breaks = c(2024, 2040, 2050, 2060, 2070, 2080, 2090, 2100)) +

# Use an escape sequence to get a new line (to match the graph from article)
labs(title = "How do UN Population projections compare to the previous\nrevision? World",
     subtitle = str_wrap("The medium population projection from the UN's World Population
                         Prospects in its 2024 publication, compared to its 2022 revision."),
     x = "Year",
     y = "Projection",
     caption = "Data Source: UN, World Population Prospects (2024)\nOurWorldinData.org/population-grow

# Modify the y-axis tick marks
# Citation: https://scales.r-lib.org/reference/unit_format.html
scale_y_continuous(labels = unit_format(unit = "Billion",
                                        scale = 1e-6)) +

# Modify the caption position using hjust (see citation below)
# Citation:
# https://www.datanovia.com/en/blog/ggplot-title-subtitle-and-caption/#change-caption-position
theme(plot.caption = element_text(hjust = 0),
      plot.title = element_text(face = "bold"))
```



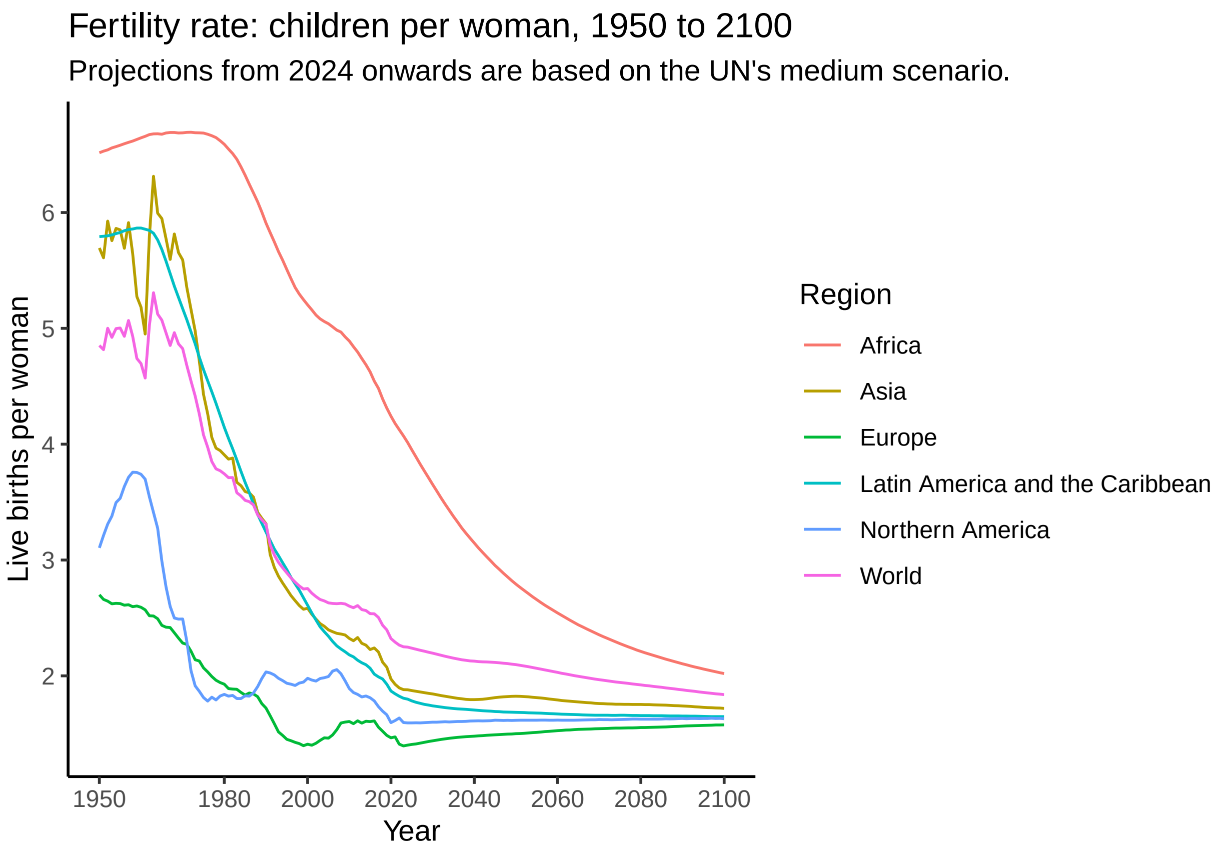
2. **Diagram 2:** Projections broken down by world and continent - Shalini

```
population_projections <- mediums |>
  select(year, region_subregion_country_area, total_pop_january_thousands) |>
  filter(region_subregion_country_area %in% c("World", "Africa", "Asia",
                                               "Europe", "Northern America",
                                               "Latin America and the Carribean"))

population_projections$facet = factor(population_projections$region_subregion_country_area,
                                      levels = c("World", "Africa", "Asia", "Europe",
                                                 "Northern America",
                                                 "Latin America and the Carribean"))


population_projections |>
  ggplot(aes(x = year,
             y = total_pop_january_thousands)) +
  geom_line() +
  facet_wrap(~facet, scales = "free",
             ncol = 2) +
  theme_minimal() +
  scale_y_continuous(labels = unit_format(unit = "B", scale = 1e-6)) +
  scale_x_continuous(breaks = c(2024, 2040, 2050, 2060, 2080, 2100)) +
  labs(title = "UN Population Projections as of 2024",
       subtitle = "Population projection from the UN World Population Prospects \nin its 2024 publicati
       x = "Year",
       y = "Population Projection (in billions)")
```



3. **Diagram 3:** Fertility rate in children/woman from 1950 - 2100 - Jonathan

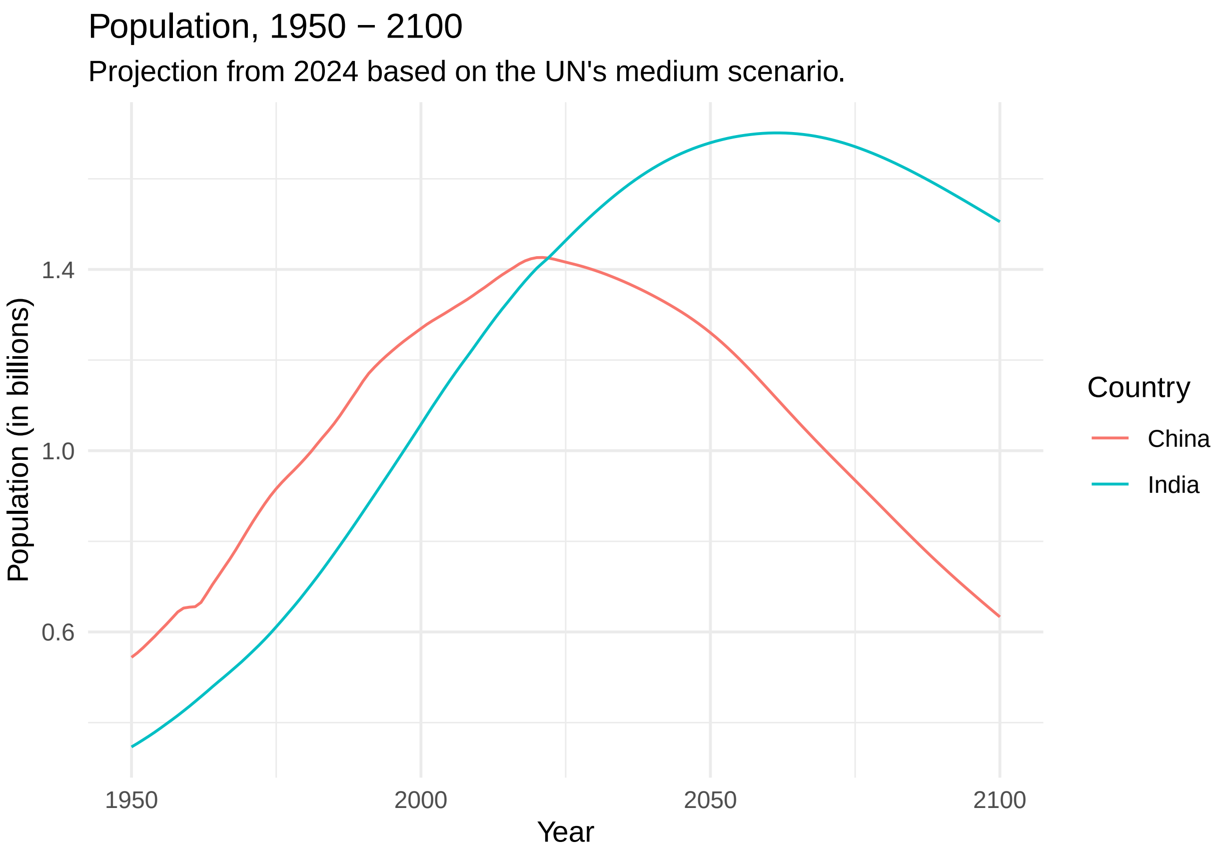
```r
estimates %>%
  rbind(mediums) %>%
  filter(region_subregion_country_area %in% c('World', 'Africa', 'Asia',
                                              'Northern America',
                                              'Latin America and the Caribbean',
                                              'Europe')) %>%
  ggplot(aes(x = year, y = total_fertility_rate_live_births_per_woman,
             color = region_subregion_country_area)) +
  geom_line() +
  scale_x_continuous(breaks = c(1950, 1980, 2000, 2020, 2040, 2060, 2080, 2100)) +
  labs(
    title = "Fertility rate: children per woman, 1950 to 2100",
    subtitle = "Projections from 2024 onwards are based on the UN's medium scenario.",
    x = "Year",
    y = "Live births per woman",
    color = "Region"
  ) +
  theme_classic()
```



4. **Diagram 4:** Population 1950 to 2100 - Alicia

```r
# Filter data for India and China from 1950 to 2100
estimates |> rbind (mediums) |>
  filter(region_subregion_country_area %in% c( "India", "China")) |>

  # Plot
ggplot(aes(x = year,
```

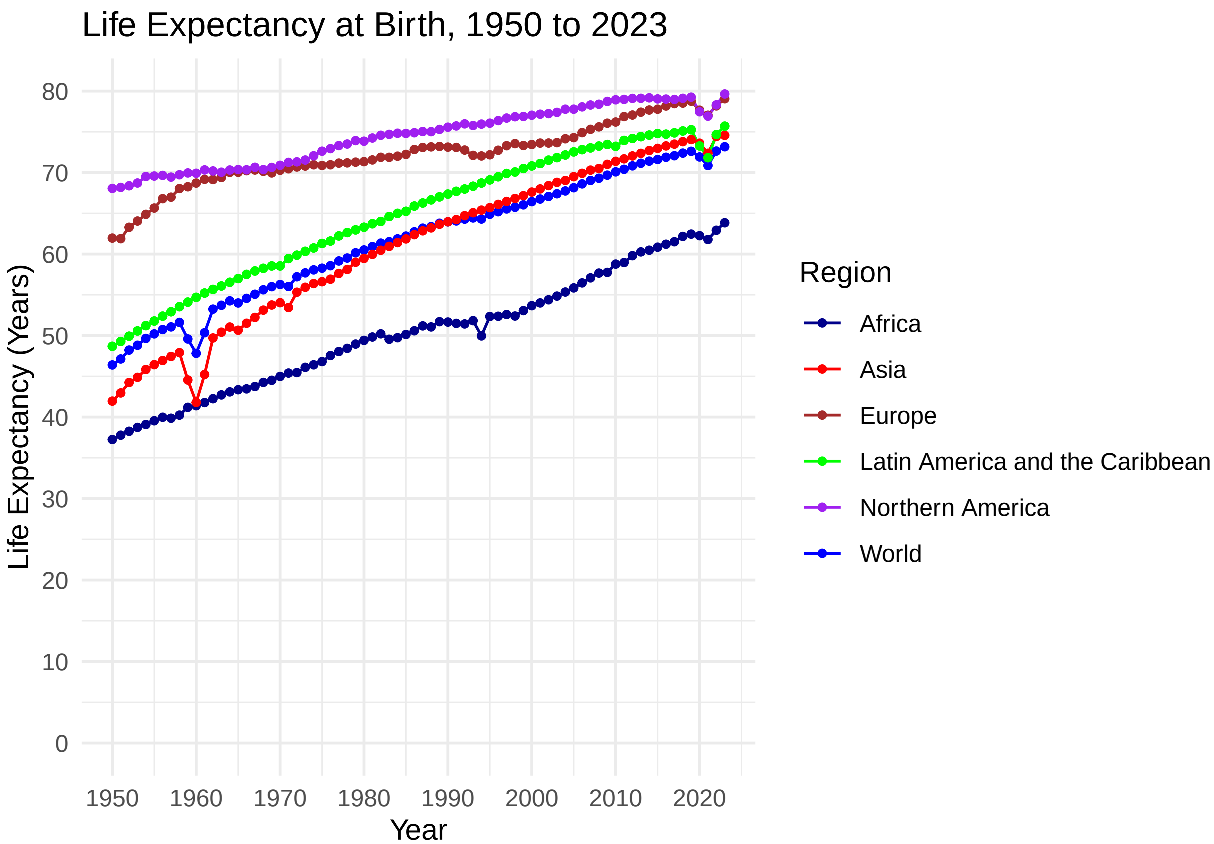
```
         y = total_pop_july_thousands / 1e6,
         color = region_subregion_country_area)) +
  geom_line() +
  labs(title = "Population, 1950 - 2100",
       subtitle = "Projection from 2024 based on the UN's medium scenario.",
       x = "Year",
       y = "Population (in billions)",
       color = "Country") +
  theme_minimal()
```



5. **Diagram 5:** Life expectancy from 1950 to 2023 - Anusha

```
life_expectancy_data <- estimates %>%
  select(year, region_subregion_country_area, total_life_expectancy_at_birth_years) %>%
  filter(region_subregion_country_area %in% c("World", "Northern America", "Europe",
                                               "Asia", "Africa",
                                               "Latin America and the Caribbean"))

ggplot(life_expectancy_data, aes(x = year,
                                 y = total_life_expectancy_at_birth_years,
                                 color = region_subregion_country_area)) +
  geom_line() +
  geom_point(size = 1) +
  labs(title = "Life Expectancy at Birth, 1950 to 2023",
       x = "Year",
       y = "Life Expectancy (Years)",
       color = "Region") +
```

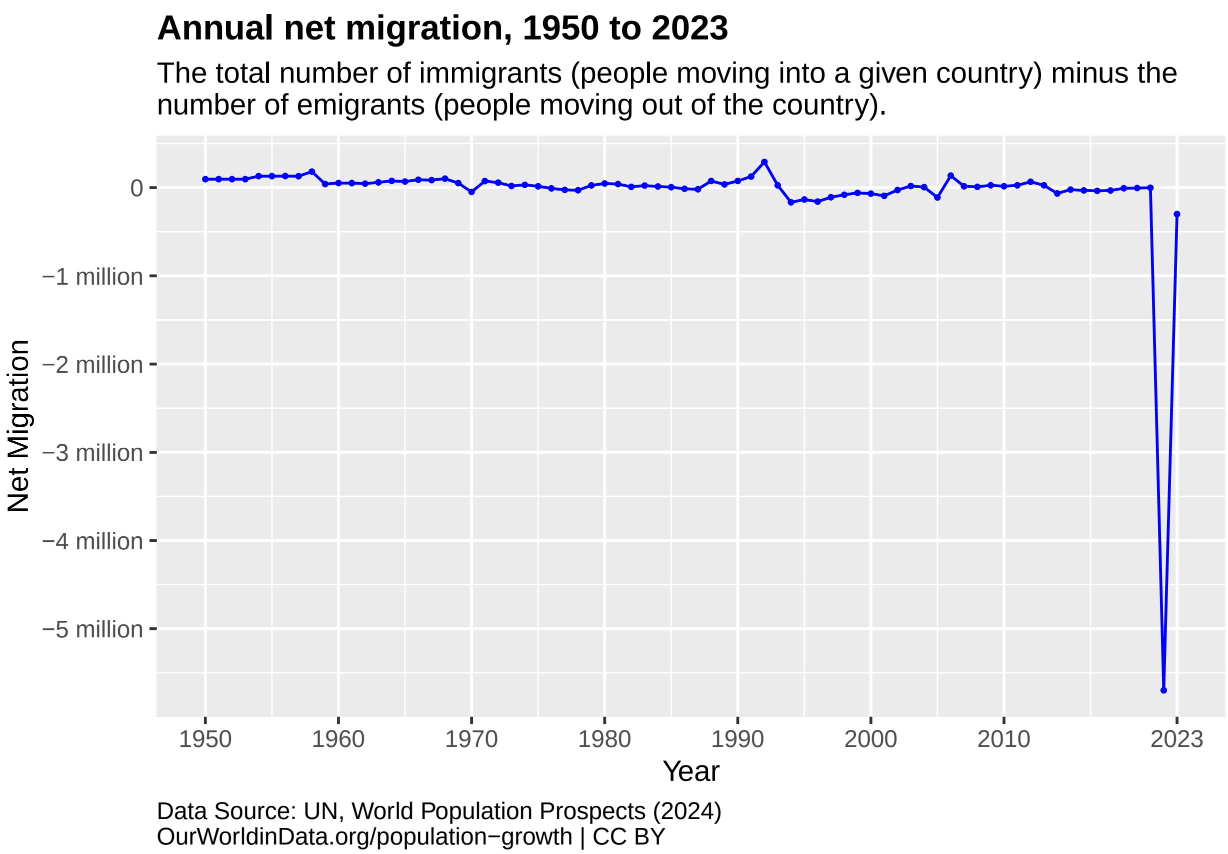
```
  scale_y_continuous(limits = c(0, 80),
                     breaks = seq(0, 80, by = 10)) +
  scale_x_continuous(limits = c(1950, 2023),
                     breaks = seq(1950, 2023, by = 10)) +
  theme_minimal() +
  scale_color_manual(values = c("World" = "blue",
                                "Northern America" = "purple",
                                "Europe" = "brown",
                                "Asia" = "red",
                                "Africa" = "darkblue",
                                "Latin America and the Caribbean" = "green"))
```



6. **Diagram 6:** Annual net migration 1950 to 2023 - Sophia Giuliani

```
estimates %>%
  select(net_num_migrants_thousands,
         year,
         region_subregion_country_area) %>%
  filter(!is.na(year)) %>%
  filter(region_subregion_country_area == "Ukraine") %>%
  filter(year <= "2023" & year >= "1950") %>%
  ggplot(mapping = aes(x = year,
                       y = net_num_migrants_thousands)) +
  geom_line(color = "blue") +
  geom_point(size = 0.5,
             color = "blue") + # Need to make the points/dots smaller
  scale_x_continuous(breaks = c(1950, 1960, 1970, 1980, 1990, 2000, 2010, 2023)) +
```

```
  scale_y_continuous(breaks = seq(-5000, 0, by = 1000),
                     labels = c("-5 million", "-4 million", "-3 million", "-2 million", "-1 million", "(
  labs(title = "Annual net migration, 1950 to 2023",
       subtitle = str_wrap("The total number of immigrants (people moving into a given country)
                           minus the number of emigrants (people moving out of the country)."),
       x = "Year",
       y = "Net Migration",

       # Use an escape sequence in the caption to get a new line
       caption = "Data Source: UN, World Population Prospects (2024)\nOurWorldinData.org/population-gro
  theme(plot.caption = element_text(hjust = 0),
        plot.title = element_text(face = "bold"))
```



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.

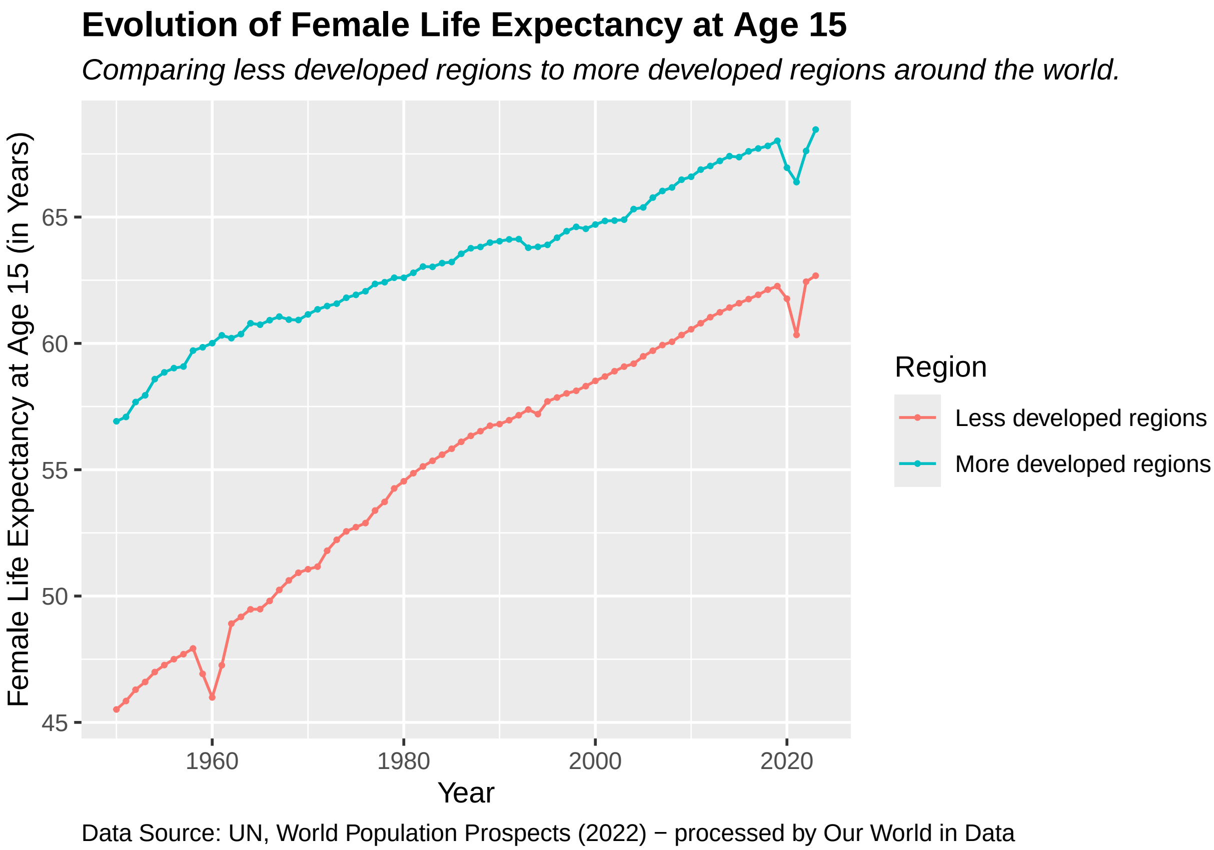
Summarize your interpretations after each chart.

1. Diagram 1: Sophia Giuliani

- **Question:** How has the female life expectancy at age 15 evovled since 1950 in less developed regions compared to more developed regions? Is there a convergence in the female life expectancy rates?
- **Interpretation of Diagram:** Since 1950, the female life expectancy (at age 15) in less developed and in more developed regions demonstrate a relatively similar trend, despite a couple differences as noted in the graph below. That is, while the female life expectancy was increasing at a relatively constant rate

from 1950 to just before 2020 in more developed regions, the upward trend in less developed regions was interrupted by a decline in the female life exptectancy for a couple years prior to 1960. However, in 1960, the female life expectancy rate in less developed countries spiked back to its previous levels and continued to increase until just before 2020. In what appears to be the year 2019, the female life expectancy rate declined in both less developed and more developed regions. Following this decline, the female life expectancy increased a couple years later. COVID-19 could be a plausable explanation for this drop that is demostrated on the graph below. In summary, there has not been a convergence in the female life expectancy rates between the two regions. However, the gap between the two regions has become more narrow overtime.

```
estimates %>%
  select(region_subregion_country_area,
         year,
         female_life_expectancy_at_age_15_years) %>%
  filter(region_subregion_country_area %in% c("More developed regions",
                                              "Less developed regions")) %>%
  ggplot(mapping = aes(x = year,
                       y = female_life_expectancy_at_age_15_years,
                       color = region_subregion_country_area)) +
  geom_line() +
  geom_point(size = 0.5) +
  scale_x_continuous(limits = c(1950, 2023)) %>%
  scale_y_continuous(breaks = c(45, 50, 55, 60, 65, 70, 75)) %>%
  labs(title = "Evolution of Female Life Expectancy at Age 15",
       subtitle = str_wrap("Comparing less developed regions to more developed regions around the world
       x = "Year",
       y = "Female Life Expectancy at Age 15 (in Years)",
       color = "Region",
       caption = "Data Source: UN, World Population Prospects (2022) - processed by Our World in Data")
  theme(plot.caption = element_text(hjust = 0),
        plot.title = element_text(face = "bold"),
        plot.subtitle = element_text(face = "italic"))
```

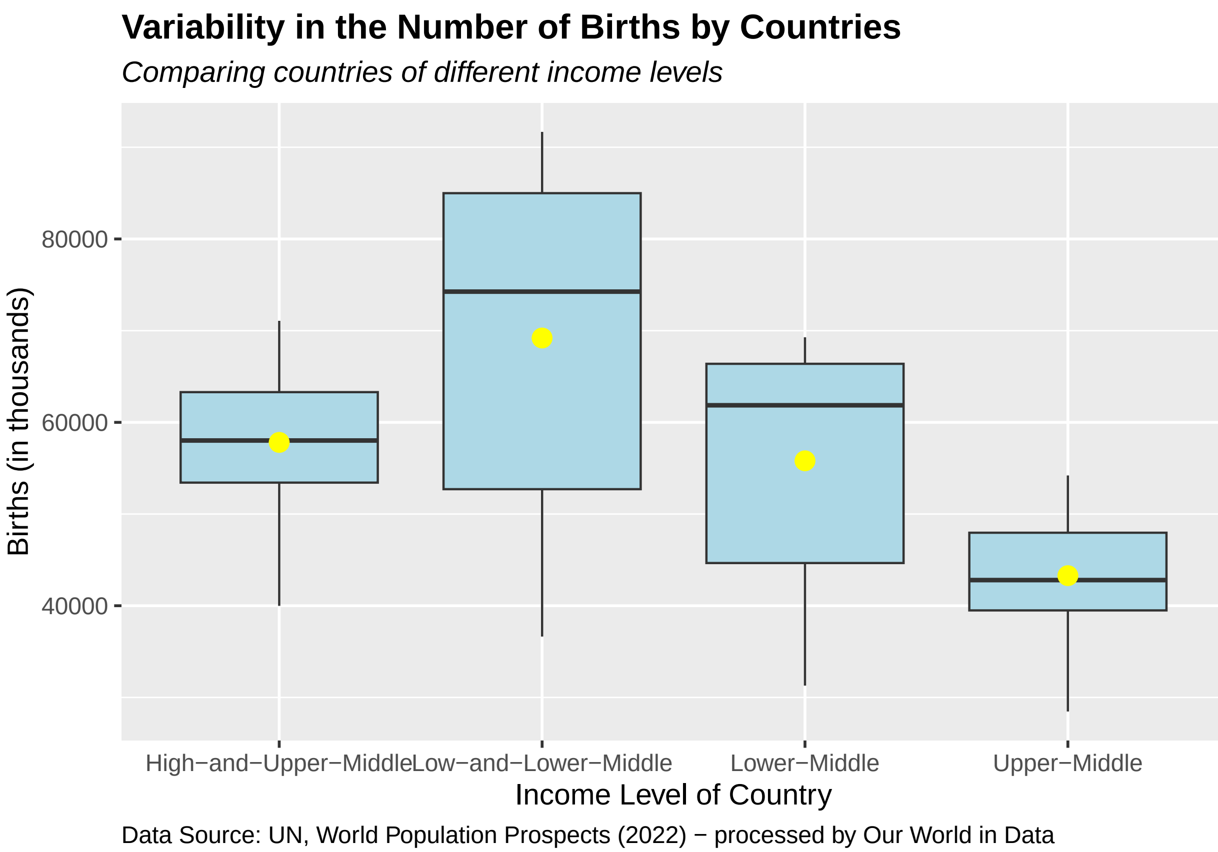


2. Diagram 2: Sophia Giuliani

- **Question:** Which income range has the largest variability in the number of births? Are countries of different income levels similar in terms of the average number of births?
- **Interpretation of Diagram:** To answer this question entailing one categorical variable and one quantitive variable, box-plots can be of great visualizations. The variability in the number of births across each of the four income levels is indicated by the IQR which represents the spread of the middle 50% of the data (citation: https://www.statology.org/box-plot-variability/). Thus, based on the diagram below, the Low-and-Lower-Middle income level countries appear to have the largest variability in the number of births, followed by countries of Lower-Middle income levels. On the other hand, Upper-Middle income level countires demonstrate the smallest variation in the number of births as the IQR is the smallest. Furthermore, we see that Low-and-Lower-Middle income level countries have the highest average number of births. It appears as though, both High-and-Upper and Lower-Middle income level countries are similar in terms of their average number of births.

```
estimates %>%
  select(region_subregion_country_area,
         births_thousands) %>%
  filter(region_subregion_country_area %in% c("Lower-middle-income countries",
                                              "Upper-middle-income countries",
                                              "High-and-upper-middle-income countries",
                                              "Low-and-Lower-middle-income countries")) %>%
  ggplot(mapping = aes(x = region_subregion_country_area,
                       y = births_thousands)) +
  geom_boxplot(linewidth = 0.4,
               fill = "lightblue") +
  stat_summary(fun = mean,
               size = 3,
```

```
              color = "yellow",
              geom = "point") +
  scale_x_discrete(labels = c("High-and-Upper-Middle",
                              "Low-and-Lower-Middle",
                              "Lower-Middle",
                              "Upper-Middle")) +
  labs(title = "Variability in the Number of Births by Countries",
       subtitle = "Comparing countries of different income levels",
       x = "Income Level of Country",
       y = "Births (in thousands)",
       caption = "Data Source: UN, World Population Prospects (2022) - processed by Our World in Data")
   theme(plot.title = element_text(face = "bold"),
         plot.subtitle = element_text(face = "italic"),
         plot.caption = element_text(hjust = 0))
```



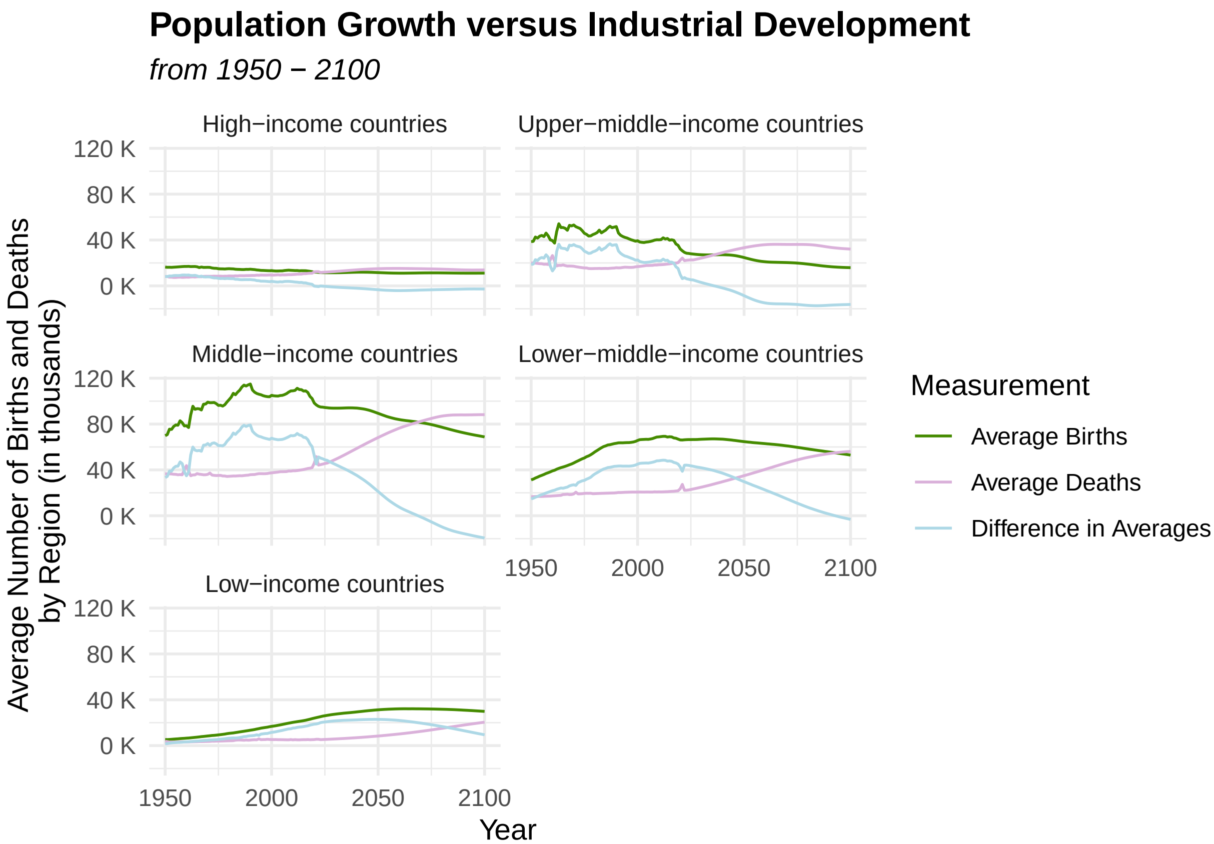
3. Diagram 3: Shalini Asokkumar

- **Question:** How is population growth affected by a country's development status?
- **Interpretation of Diagram:** The least shocking of the results is shown in the high income countries bracket, where the net births and deaths remained relatively stable except for a few shocks here and there. These can likely be associated with times of economic turmoil and income struggles. However, it is important to note, that recently these countries populations are predicted to be shrinking at an increasing rate. Many economists are calling this a cause for concern as it means there may be fewer people able to take on existing jobs. Across all 5 graphs, we see, that contrary to increasing life expectancy, people in all countries are, on average, expected to have fewer children. This is especially interesting to note in the Lower Middle Income Countries and the Low income countries, because their populations have been growing at an increasing rate for since before the 50s. Likely, this was due to

cultural practices of having large families. So what can we attribute the decrease to? Recently, in higher income countries, as the costs of living increase, people have intentionally opted to have smaller families, and having no kids is also becoming a more popular idea. It could also be due to resource constraints. Parents want to give their kids the best lives possible, and may not want to have kids if they cannot ensure that possibility. The death rates in the lower income countries are can be explained by increased access to resources, such as healthcare, resulting in a demographic transition with lower fertility rates.

```
est_and_med <- rbind(estimates, mediums)

est_and_med$region_subregion_country_area = factor(est_and_med$region_subregion_country_area,
                                      levels = c("High-income countries",
                                                 "Upper-middle-income countries",
                                                 "Middle-income countries",
                                                 "Lower-middle-income countries",
                                                 "Low-income countries"))

est_and_med |>
  filter(region_subregion_country_area %in% c("High-income countries",
                                                 "Upper-middle-income countries",
                                                 "Middle-income countries",
                                                 "Lower-middle-income countries",
                                                 "Low-income countries")) |>
  group_by(region_subregion_country_area, year) |>
  summarise(`Average Births` = mean(births_thousands, na.rm = T),
            `Average Deaths` = mean(total_deaths_thousands, na.rm = T),
            `Difference in Averages` = `Average Births` - `Average Deaths`) |>
  pivot_longer(cols = c("Average Births", "Average Deaths", "Difference in Averages"),
               names_to = "rate_type", values_to = "averages") |>
  ggplot(aes(x = year, y = averages, color = rate_type)) +
  scale_color_manual(values = c('chartreuse4','#DAB1DA', "lightblue"))+
  geom_line() +
  scale_y_continuous(labels = unit_format(unit = "K", scale = 1e-3)) +
  facet_wrap(~region_subregion_country_area, ncol = 2) +
  labs(title = "Population Growth versus Industrial Development",
       subtitle = "from 1950 - 2100",
       x = "Year",
       y = "Average Number of Births and Deaths \nby Region (in thousands)",
       color = "Measurement") +
  theme_minimal() +
  theme(plot.title = element_text(face = "bold"),
        plot.subtitle = element_text(face = "italic"))
```

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

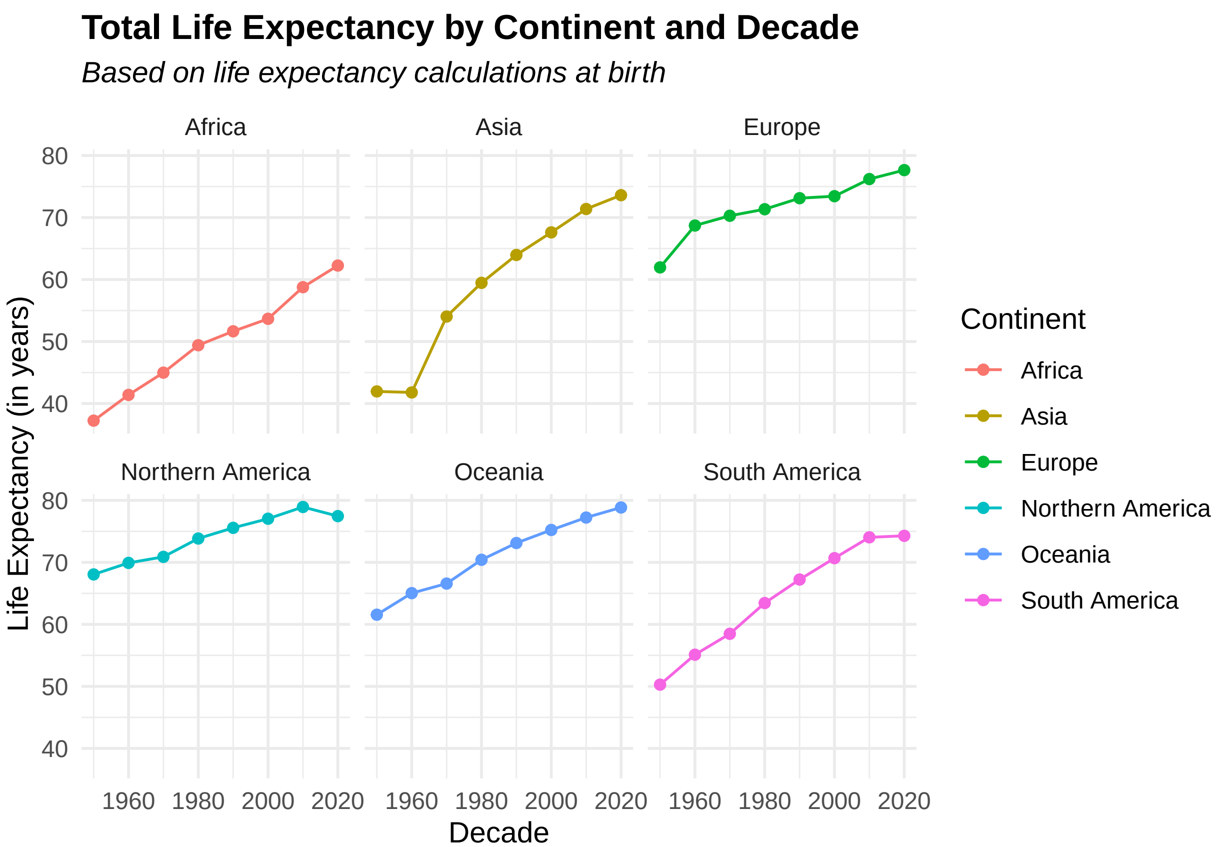


4. Diagram 4: Shalini Asokkumar

- **Question:** How has total life expectancy at birth changed by continent at the beginning of each decade?
- **Interpretation of Diagram:** Before creating this model, I assumed that North America, Oceania, and Europe were going to consistently have the highest life expectancy rates. This can be supported by how these predominantly white areas have institutionalized a higher standard of living than coutnries in the Global South. They were able to build their systems based on colonialism and, now, continue to benefit from importing goods made with manual labor. They also have more robust and established health systems with government support. I was surprised to see that Asia and South America showed such promising and significant improvements in life expectancy; however, I think this can be reconciled with the growing prominence of the BIRCs countries in the global economy, or Brazil, Russia, India, and China. They are now seeing increased investment in their commodities from countries in the Global North, and foreign interest in supporting their economic development. As a result of being given access to increased financial resources, it makes sense that citizens are living longer, and presumably, better quality lives.

```
estimates |> filter(region_subregion_country_area %in%
                      c("Northern America", "South America", "Asia",
                        "Oceania", "Europe", "Africa")) |>
  filter(year == 1950 | year == 1960 |
         year == 1970 | year == 1980 |
         year == 1990 | year == 2000 |
         year == 2010 | year == 2020) |>
  group_by(region_subregion_country_area, year) |>
  mutate(mean_life_exp = mean(total_life_expectancy_at_birth_years)) |>
  select(region_subregion_country_area, year, mean_life_exp) |>
  ggplot(aes(x = year, y = mean_life_exp, color = region_subregion_country_area)) +
```

```
geom_line() +
geom_point() +
facet_wrap(~region_subregion_country_area) +
labs(title = "Total Life Expectancy by Continent and Decade",
     subtitle = "Based on life expectancy calculations at birth",
     x = "Decade",
     y = "Life Expectancy (in years)",
     color = "Continent") +
theme_minimal() +
theme(plot.title = element_text(face = "bold"),
        plot.subtitle = element_text(face = "italic"))
```

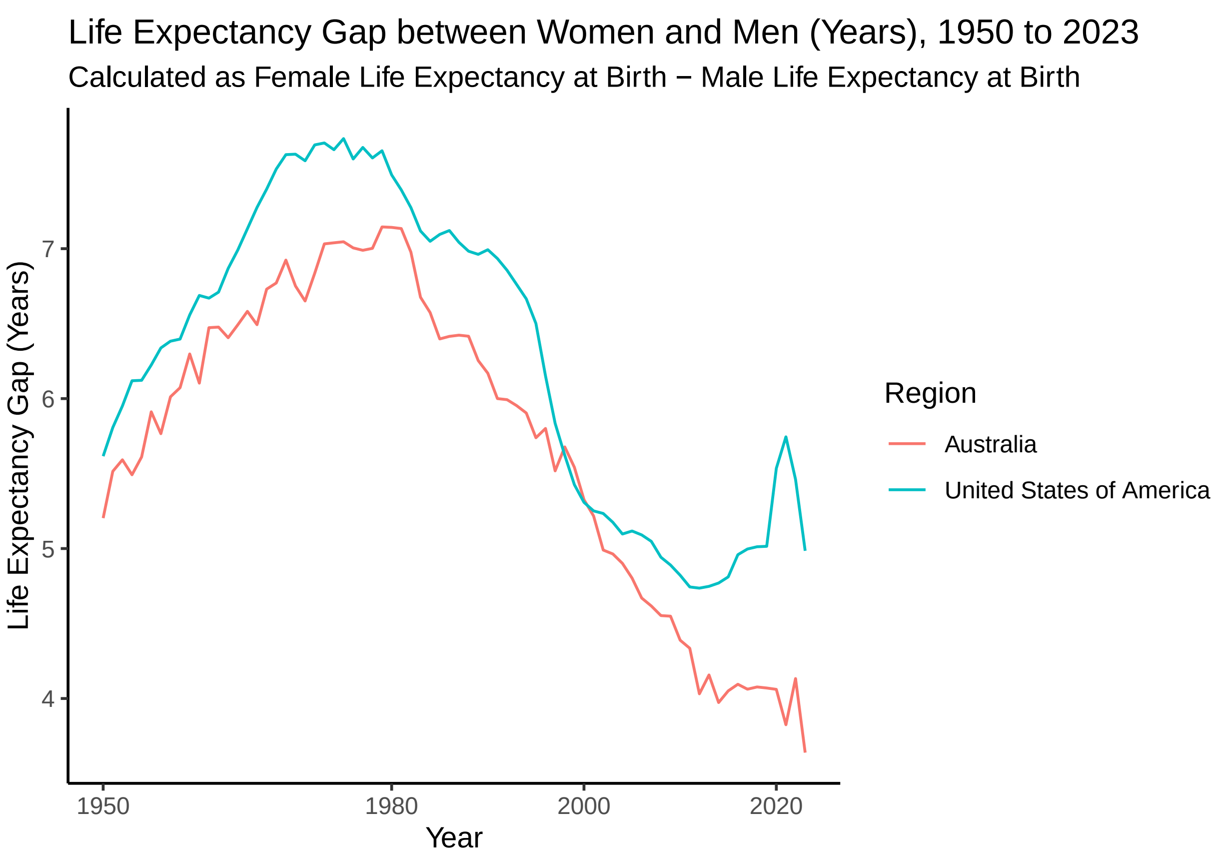


5. Diagram 5: Jonathan Sarasa

- **Question:** How has the Female/Male Life Expectancy Gap changed between the United States and Australia?
  - **Interpretation of Diagram:** In both the United States and Australia populations, females have consistently longer lifespans than men. In this diagram, the difference varies between 5 years - 8 years for the United States and between 3.7 - 7 years for Australia (between 1950-2023). Both the USA and Australia show similar trends in terms of female/male life expectancy gap changes. In 1950, the life expectancy gap was between 5-6 years for both countries before shooting up to its largest gap in the 1980s then declining into the 21st century. The life expectancy gap between women and men has been larger in the USA, except during a brief period around 1999 - when Australia had a larger gap than the USA. Covid-19 affected both of these charts, as it appears that Covid caused in the USA a drastic spike in male mortalility which was not matched my female rates of increased mortality. This caused an increase in the gap between male and female life expectancy at birth in 2020. Interestingly, around 2020, Australia shows a decrease in the gap

between male and female life expectancy and then a small spike upwards around 2020. Covid-19 produced two different patterns in the female/male life expectancy gap in USA and Australia. This may be because Covid-19 did not cause as many deaths as much in Australia as it did in the US. As of 2024, Australia has 24,414 deaths from Covid while the USA has 1,219,487 deaths from Covid (https://www.worldometers.info/coronavirus).

```
estimates %>%
  filter(region_subregion_country_area %in% c('United States of America', 'Australia')) %>%
  ggplot(aes(x = year,
             y = female_life_expectancy_at_birth_years - male_life_expectancy_at_birth_years,
             color = region_subregion_country_area)) +
  geom_line() +
  scale_x_continuous(breaks = c(1950, 1980, 2000, 2020)) +
  labs(
    title = "Life Expectancy Gap between Women and Men (Years), 1950 to 2023",
    subtitle = "Calculated as Female Life Expectancy at Birth - Male Life Expectancy at Birth",
    x = "Year",
    y = "Life Expectancy Gap (Years)",
    color = "Region"
  ) +
  theme_classic()
```



6. Diagram 6: Jonathan Sarasa

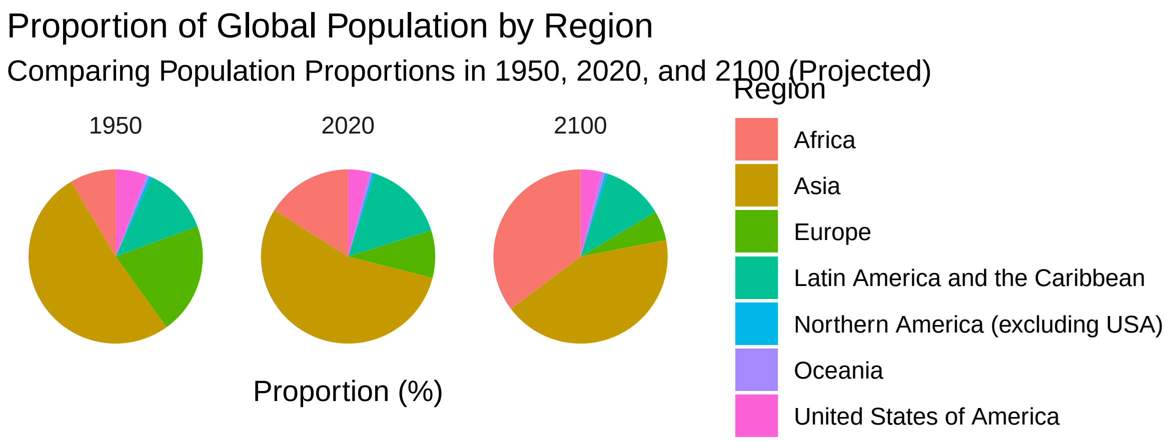
- **Question:** Comparing 1950 vs 2020 vs 2100, what proportion of global population growth did each region of the world contribute?
  - **Interpretation of Diagram:** This graphs shows shares of global population by continent in the year 1950, 2020, and 2100 (predicted). These graphs shows the massive population of Asia,

relative to the other continents/regions of the World. I separated out the USA from the rest of Northern America, but the USA is the majority of the North American population. This graph shows the decline of USA and Europe as major shares of the world population, with both portions shrinking over the years. In the year of 2100, the population of Asia is almost matched by Africa. African countries are expected to have massive increases in populations as development continues, infant mortality decreases, and life expectancy increases significantly. The 2100 projection shows an increased world population which is majority African and Asian. These graphs also highlight how small the populations of USA and Europe are compared to the rest of the world, which stands in contrast to their economic power and colonialism over other world regions.

```r
#Regions: USA, North America - USA, Latin America & Caribbean, Europe, Asia, Africa, Oceania
graph6data <- estimates %>%
  rbind(mediums) %>%
  filter(region_subregion_country_area %in% c("United States of America",
                                              "Northern America",
                                              "Latin America and the Caribbean",
                                              "Europe", "Asia", "Africa", "Oceania")) %>%
  filter(year %in% c(1950, 2020, 2100)) %>%
  group_by(year, region_subregion_country_area) %>%
  summarise(total_population = sum(total_pop_january_thousands, na.rm = TRUE))
```

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

```r
# Remove USA from North America population
graph6data <- graph6data %>%
  mutate(total_population =
         ifelse(region_subregion_country_area == "Northern America",
                total_population - sum(total_population[region_subregion_country_area ==
                                                         "United States of America"], na.rm = TRUE),
                total_population)) %>%
  group_by(year) %>%
  mutate(global_population = sum(total_population, na.rm = TRUE)) %>%
  mutate(proportion = total_population / global_population * 100) %>%
  mutate(region_subregion_country_area =
         ifelse(region_subregion_country_area == "Northern America", "Northern America (excluding USA)
                region_subregion_country_area))

# pie chart
graph6data %>%
  ggplot(aes(x = "", y = proportion, fill = region_subregion_country_area)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar("y", start = 0) +
  facet_wrap(~year) +
  labs(
    title = "Proportion of Global Population by Region",
    subtitle = "Comparing Population Proportions in 1950, 2020, and 2100 (Projected)",
    fill = "Region",
    x = "",
    y = "Proportion (%)"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_blank(),
  axis.ticks = element_blank(),
  panel.grid = element_blank())
```

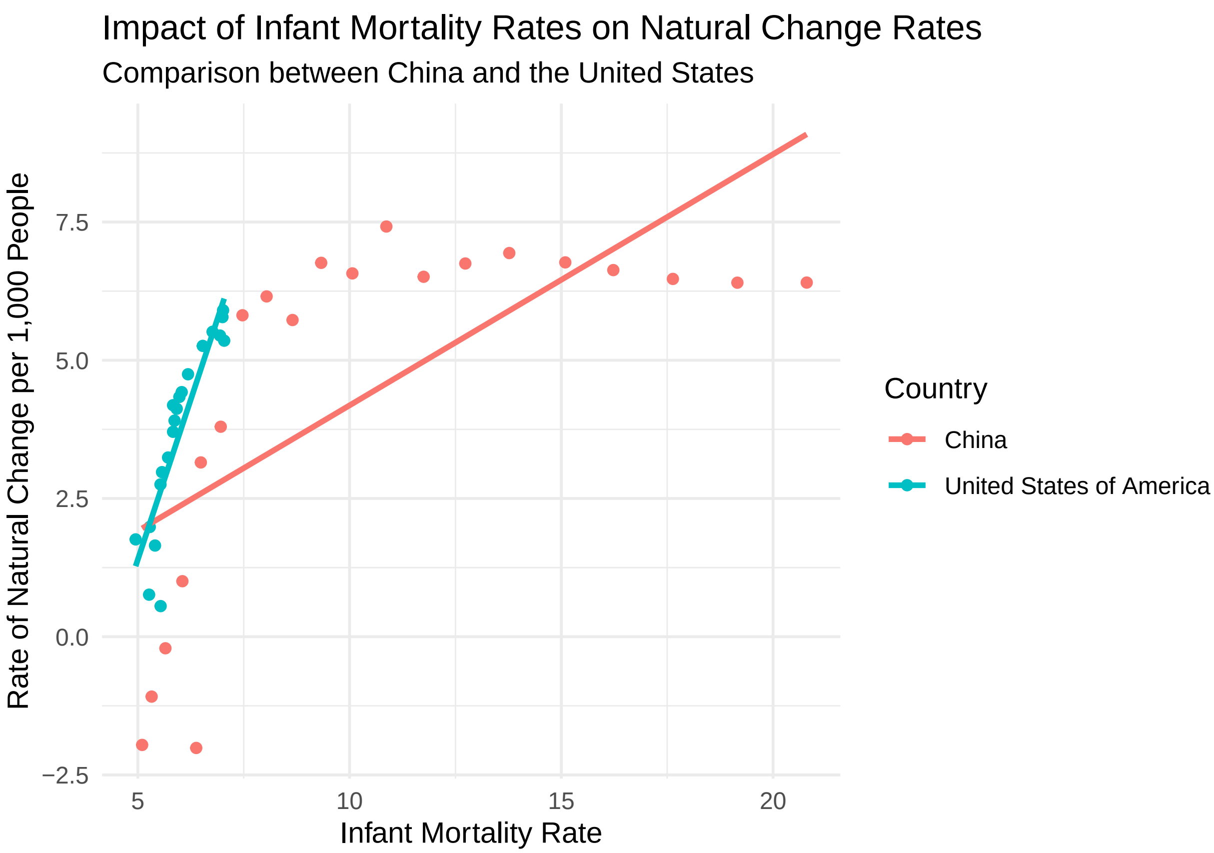


7. Diagram 7 Alicia Zhou

- **Question:** How do the Infant Mortality Rates of China and the United States affect their natural change rates over the past 2 decades?
- **Interpretation:** The data for China spans a wide range of infant mortality rates compared to the U.S. This indicates that there have been significant variations in healthcare access, socioeconomic conditions, or other factors affecting infant mortality over time or across regions. But the variability in natural change rate is more moderate in China compared to the U.S. This suggests that other factors have dampened the impact of infant mortaility rate on natural change rates. For the U.S, the low variability in infant mortality reflects consistent healthcare outcomes across regions and time. On the other hand, the high variability in natural changes suggest that other factors besides infant mortality rates are influencing natural change rates. The relatively steeper positive slope for the U.S. highlights a more sensitive elationship between infant mortality and natural change rates. Small changes in infant mortality seem to correspond to larger changes in the natural change rate, possibly due to the demographic dynamics of a more developed population. The flater slope for China indicates that the historical one-child policy may have diminished the impact.

```r
# Set the time frame
current_year <-2024
start_year <- current_year - 20

# Get the necessary data
estimates |> rbind(mediums) |>
  filter (region_subregion_country_area %in% c("China", "United States of America"),
          year >= start_year, year<= current_year) |> # Filter out the rows
  select (region_subregion_country_area, year,
          rate_of_natural_change_per_1000,
          infant_morality_rate_infant_deaths_per_1000_births) |>

# Plot
ggplot(aes(x = infant_morality_rate_infant_deaths_per_1000_births,
           y = rate_of_natural_change_per_1000,
           color = region_subregion_country_area)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) + # Smooth the points to see trend
  labs(title = "Impact of Infant Mortality Rates on Natural Change Rates",
       subtitle = "Comparison between China and the United States",
       x = "Infant Mortality Rate",
       y = "Rate of Natural Change per 1,000 People",
       color = "Country") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

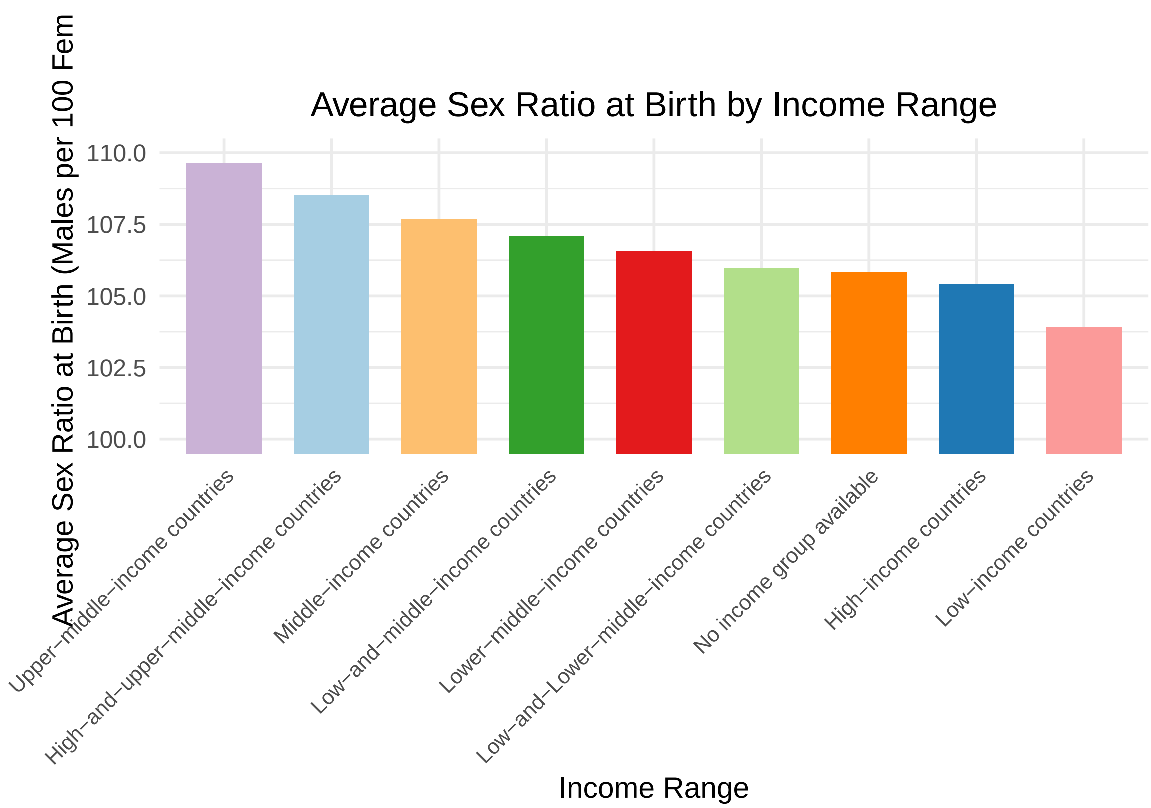


8. Diagram 8 Alicia Zhou

- **Question:** How does the income range of different countries correlate with the sex ratio at birth, and are there observable differences?
- **Interpretation:** The income group with the highest sex ratio is upper-middle-income countries (around 110 males per 100 females), the income group with the lowest sex ratio is low-income countries (around 104 males per 100 females). In general, lower the income, lower the average sex ratio at birth. However, surprisingly, high-income countries have the second lowest sex ratio at birth, reflecting relatively balanced demographic pattern. The elevated ratios in upper-middle-income and high-income countries might reflect gender preferences, particularly in regions where son preference exists. It could also indicate the use of medical technologies like prenatal sex determination followed by selective practices. Low-income countries may reflect minimal external intervention in the natural sex ratio due to limited access to such technologies or stronger adherence to natural birth outcomes.

```r
# Set Time Frame
current_year <-2023
start_year <- current_year - 30

# Group the Income Ranges and Find Average Sex Ratio at Birth
estimates |>
  filter(region_subregion_country_area %in% c("High-and-upper-middle-income countries",
                                               "Low-and-Lower-middle-income countries",
                                               "High-income countries",
                                               "Low-and-middle-income countries",
                                               "Middle-income countries",
                                               "Upper-middle-income countries",
```

```r
                                                "Lower-middle-income countries",
                                                "Low-income countries",
                                                "No income group available"),
         year >= start_year, year<= current_year) |>
  select(region_subregion_country_area,
         sex_ratio_at_birth_males_per_100_female_births, year) |>
  group_by(region_subregion_country_area) |>
  summarize(AverageSexRatio = mean(sex_ratio_at_birth_males_per_100_female_births, na.rm = TRUE,
                                   count = n(), .group = 'drop')) |>

# Plot, bar chart shows distribution
# Order the bars by decreasng average sex ratio to better visualize the highest and the lowest
ggplot(aes(x = reorder(region_subregion_country_area, - AverageSexRatio),
           y = AverageSexRatio, fill = region_subregion_country_area)) +
  geom_bar(stat = "identity", width = 0.7, show.legend = FALSE) +
  scale_fill_brewer(palette = "Paired") +
  labs(title = "Average Sex Ratio at Birth by Income Range",
       x = "Income Range",
       y = "Average Sex Ratio at Birth (Males per 100 Female)") +
  theme_minimal() +
  coord_cartesian(ylim = c(100, 110)) +
  theme(axis.text.x = element_text(angle = 45,
                                   hjust = 1, # create space for labels
                                   size = 8),
        plot.title = element_text(hjust = 0.5),
        plot.margin = margin(1,1,1,1, "cm"))
```

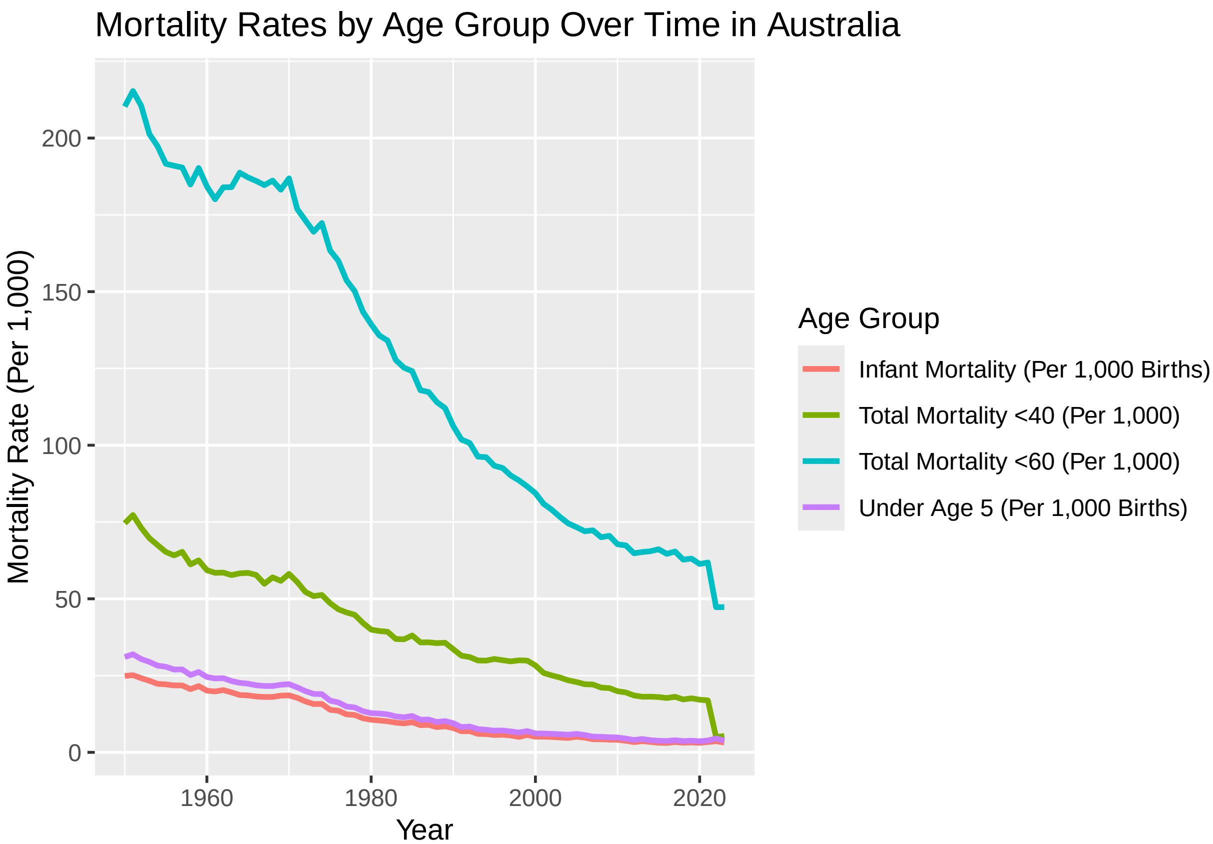


9. Diagram 9 Anusha Chinthamaduka

- **Question:** How have mortality rates across different age groups changed over time in Australia?
- **Interpretation**: This graph illustrates mortality rates in Australia (from 1950 to 2020) by various age groups (infant mortality, total mortality under age 40, total mortality under age 60, and under age 5). In the graph, all age groups show a significant decline in mortality rates over time, which can be attributed to the rapid improvements in health care and living conditions. The mortality rates of those under 60 showed the most significant and steepest decline. This could be due to the increased access in healthcare that occurred in the 1980s due to the introduction of Australia's Universal Health Care system. In addition, infant mortality rates and under 5 mortality rates seem to converge close to zero around 2020. These declines are most likely a result from advancements in neonatal care and widespread vaccination programs. Overall by 2020, mortality rates across all age groups have decreased dramatically, emphasizing Australia's progress in public health and the near-zero rates for infant and under-5 mortality, showcase the significant advancement in reducing preventable deaths in Australia.

```
australia_data <- estimates %>%
  filter(region_subregion_country_area == "Australia") %>%
  select(year,
         infant_morality_rate_infant_deaths_per_1000_births,
         "deaths_under_age_5_per_1,000_live_births",
         total_male_mortality_before_age_40_per_1000_births,
         total_mortality_before_age_60_per_1000_births) %>%
  pivot_longer(cols = starts_with("infant_morality_rate"):starts_with("total_mortality"),
               names_to = "age_group",
               values_to = "mortality_rate") %>%
  mutate(age_group = case_when(
    age_group == "infant_morality_rate_infant_deaths_per_1000_births" ~ "Infant Mortality (Per 1,000 Bi
    age_group == "deaths_under_age_5_per_1,000_live_births" ~ "Under Age 5 (Per 1,000 Births)",
    age_group == "total_male_mortality_before_age_40_per_1000_births" ~ "Total Mortality <40 (Per 1,000
    age_group == "total_mortality_before_age_60_per_1000_births" ~ "Total Mortality <60 (Per 1,000)",
    TRUE ~ age_group
  ))

ggplot(australia_data, aes(x = year,
                           y = mortality_rate,
                           color = age_group,
                           group = age_group)) +
  geom_line(size = 1) +
  labs(title = "Mortality Rates by Age Group Over Time in Australia",
       x = "Year",
       y = "Mortality Rate (Per 1,000)",
       color = "Age Group")
```

```
## 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.
```



10. Diagram 10 Anusha Chinthamaduka

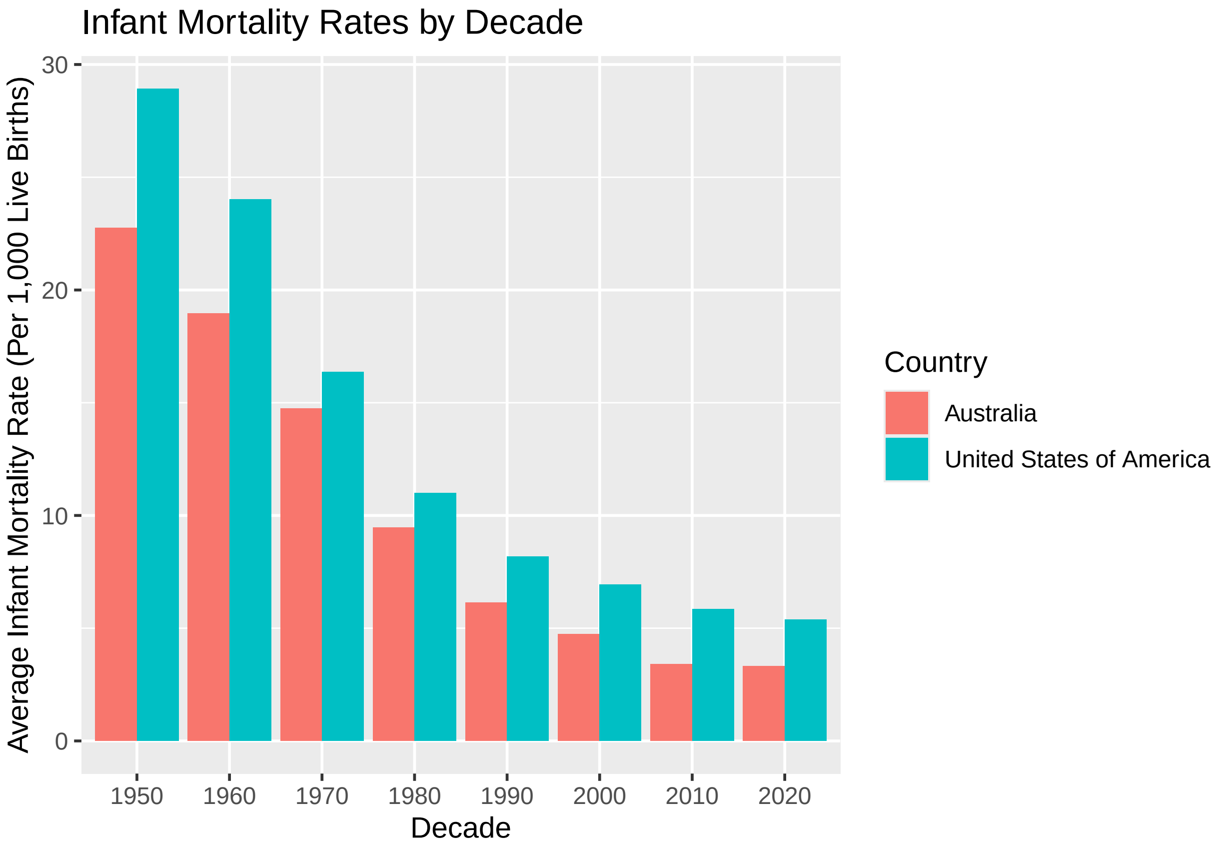
- **Question:** How do the infant mortality rates in Australia and the United States compare across different decades?
- **Interpretation** This bar chart compares the average infant mortality rates (per 1,000 live births) in Australia and the United States across decades, spanning from 1950 to 2020. According to the graph, it appears that both Australia and the United States experienced a significant decline in infant mortality rates over the decades. The United States seems to have a consistently higher infant mortality rate than Australia across all decades but it also had a more significant decline. The introduction of Medicaid in 1965 could be a contributor to this steep decline. By 2020, it seems like both countries achieved a low infant mortality rate of approximately 2–4 per 1,000 live births which reflects significant advancements in public health initiatives.

```
estimates <- estimates %>%
  mutate(decade = floor(year / 10) * 10)

filtered_data <- estimates %>%
  filter(region_subregion_country_area %in% c("Australia",
                                               "United States of America")) %>%
  group_by(decade, region_subregion_country_area) %>%
  summarise(avg_infant_mortality_rate =
              mean(infant_morality_rate_infant_deaths_per_1000_births, na.rm = TRUE))
```

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

```
ggplot(filtered_data, aes(x = factor(decade),
                          y = avg_infant_mortality_rate,
                          fill = region_subregion_country_area)) +
```

```r
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Infant Mortality Rates by Decade",
       x = "Decade",
       y = "Average Infant Mortality Rate (Per 1,000 Live Births)",
       fill = "Country")
```

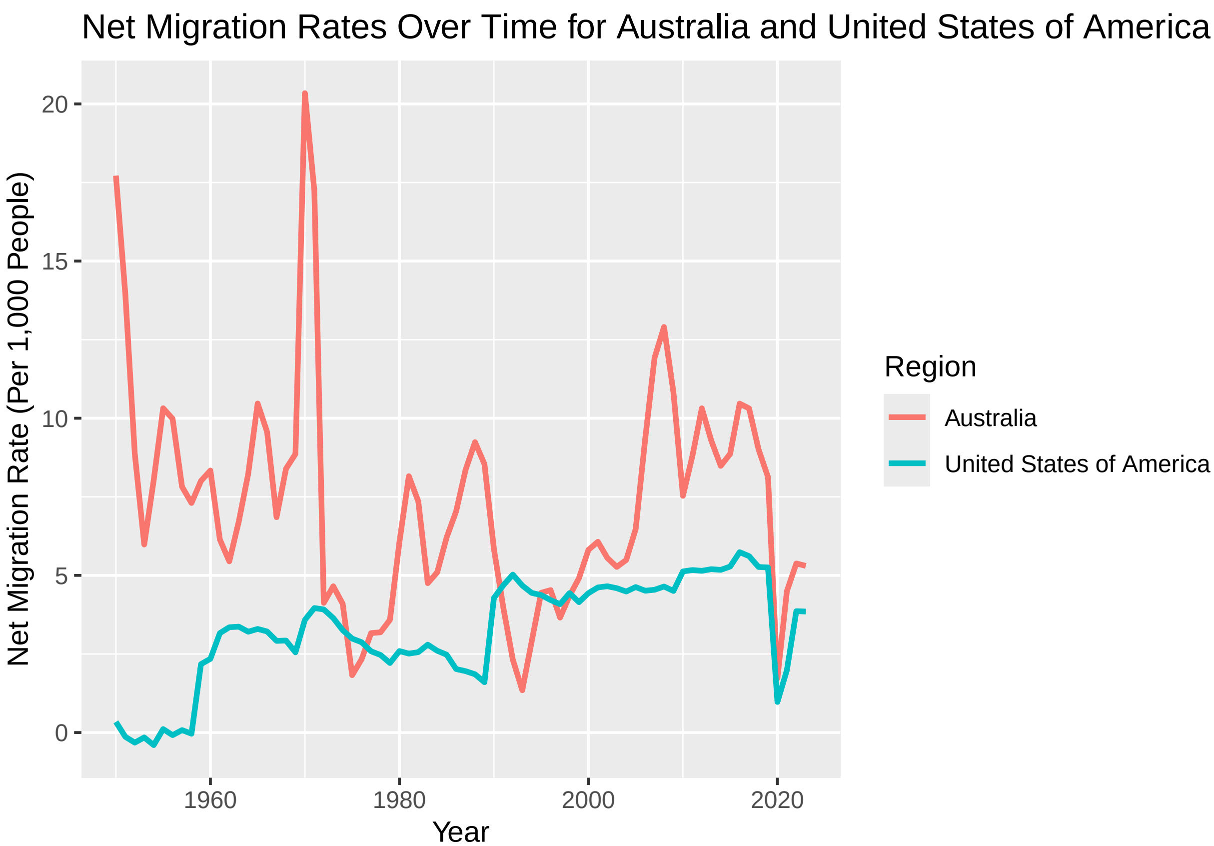


11. Diagram 11 Anusha Chinthamaduka

- **Question:** How have net migration rates changed over time in Australia vs United States?
- **Interpretation:** This line chart displays net migration rates (per 1,000 people) for Australia and the United States from 1950 to 2020. According to the chart, Australia's net migration rates appear to be consistently higher and more variable than the United States over time. This could be due to Australia's more active pro-immigration policies than the US as well as Australia's reliance on immigration for economic growth (specifically to address labor shortages by immigrating skilled migrants). There also appears to be a peak net migration rate in Australia around 1970. This could be due to the "End of the White Australia Policy", which made it easier for individuals of other races to immigrate to Australia. There is also a sharp decline in immigration for both countries in 2020, which is a direct result of the COVID-19 pandemic and the strict immigration laws that were present during this time.

```r
filtered_data <- estimates %>%
  filter(region_subregion_country_area %in% c("Australia", "United States of America")) %>%
  select(year, region_subregion_country_area, net_migration_rate_per_1000)

ggplot(filtered_data, aes(x = year,
                          y = net_migration_rate_per_1000,
                          color = region_subregion_country_area,
                          group = region_subregion_country_area)) +
  geom_line(size = 1) +
```

```r
  labs(title = "Net Migration Rates Over Time for Australia and United States of America",
       x = "Year",
       y = "Net Migration Rate (Per 1,000 People)",
       color = "Region")
```

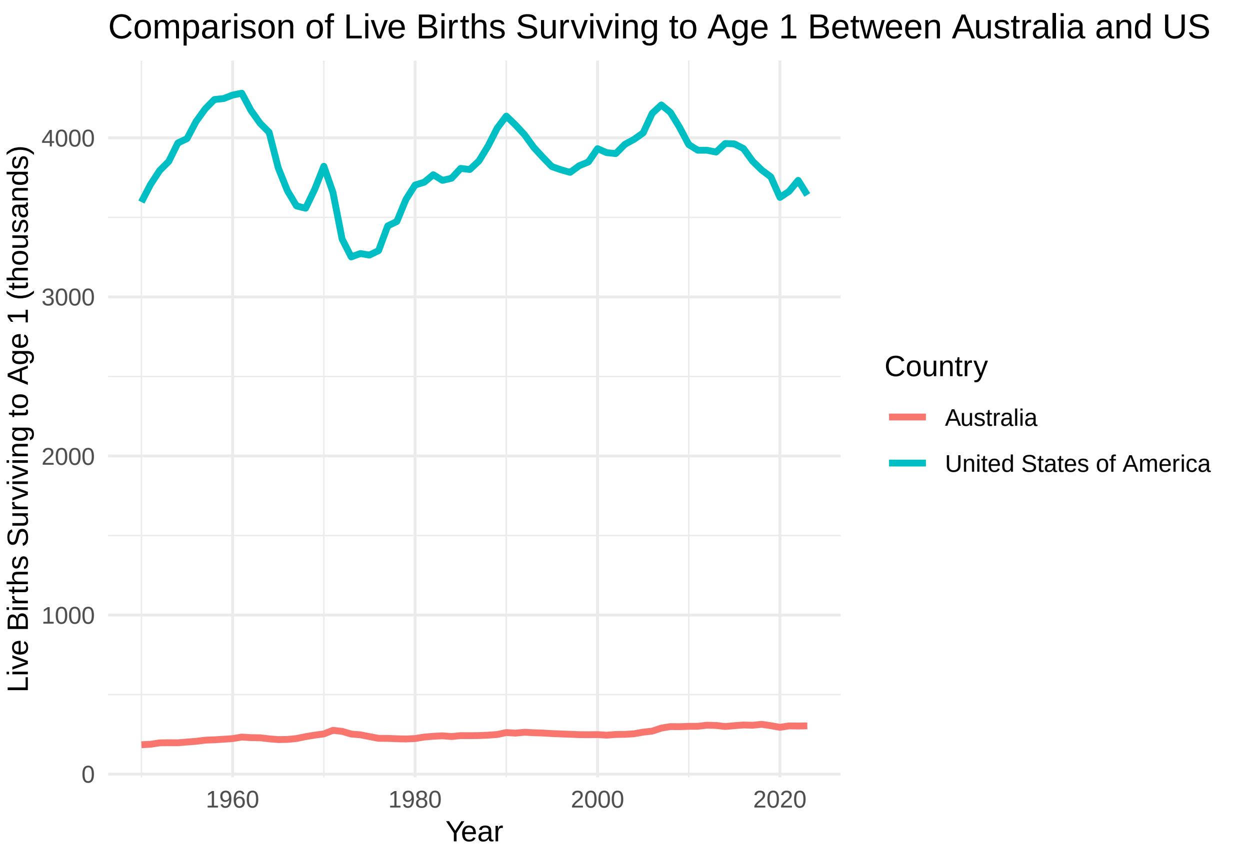


12. Diagram 12 Alicia Zhou

- **Question:**
- **Interpretation:** The number of live births surviving to age 1 in the U.S. is significantly higher than in Australia throughout the time period. This reflects the much larger population in the U.S. The U.S. line shows an increase during the post-World War II baby boom (1940s–1960s) and a noticeable decline in the 1970s, which aligns with demographic changes such as lower fertility rates following the baby boom. The numbers stabilize after the 1980s but exhibit a slight decline in the 2000s, reflecting trends in lower birth rates or other demographic changes. The number of live births surviving to age 1 in Australia remains relatively constant. The trend shows a gradual increase over time but without significant fluctuations. This steady trend reflects Australia's relatively stable fertility rates and improvements in healthcare access over the years.

```r
# Get necessary data
estimates |>
  filter(region_subregion_country_area %in% c("Australia", "United States of America")) |>
  select(region_subregion_country_area, year, live_births_surviving_to_age_1_thousands) |>

# Plot, line chart better for observing trends
  ggplot(aes(x = year,
             y = live_births_surviving_to_age_1_thousands,
             group = region_subregion_country_area,
             color = region_subregion_country_area)) +
```

```r
geom_line(size = 1.2) +
labs(title = "Comparison of Live Births Surviving to Age 1 Between Australia and US",
     x = "Year",
     y = "Live Births Surviving to Age 1 (thousands)",
     color = "Country") +
theme_minimal()
```



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.

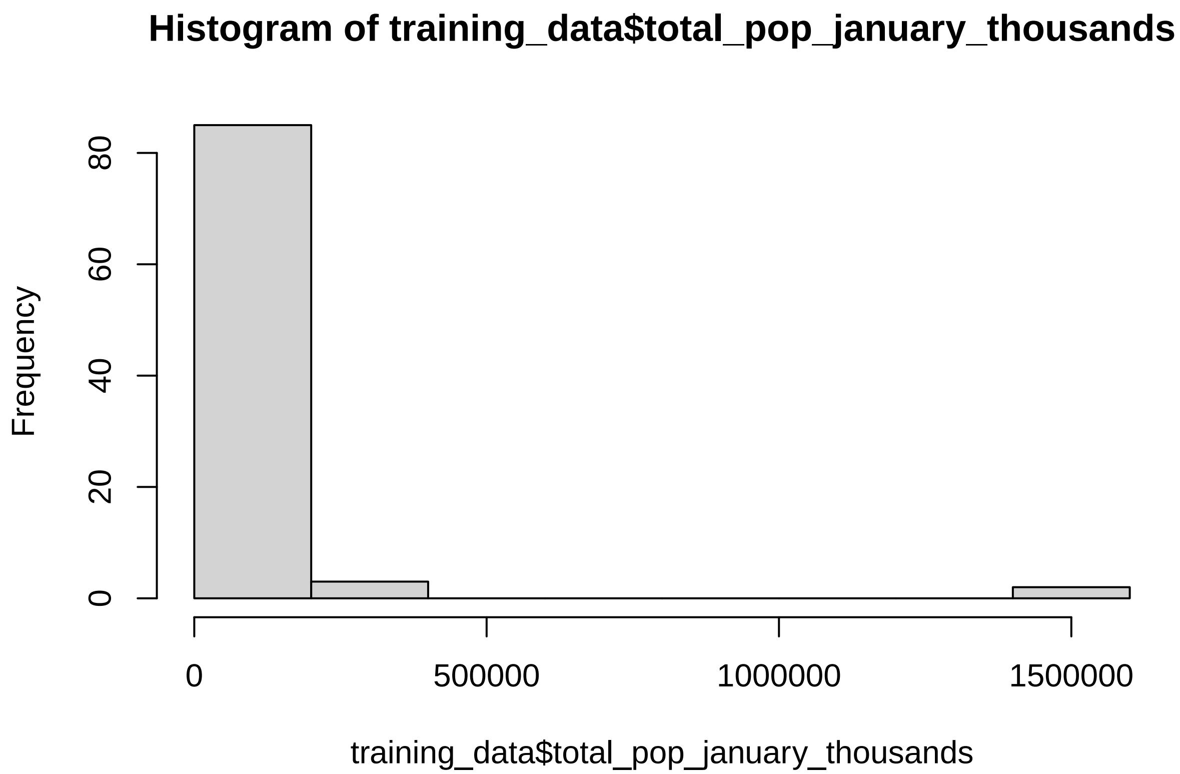
How significantly do your population projections diverge from those of the United Nations? Provide a comparison of the two. If you choose a different projection for which there is no UN data to compare with, then this comparison is not required.

```r
# ~80:20 Training to Predicting Ratio
# Training based only on 2023 numbers.
# We're predicting the 2100 number for future population of countries using
# population, fertility, mortality, & migration data.
# 237 Unique countries/areas
# Train on 190, predict the future populations of the last 47
```

```r
unique_countries <- estimates %>%
  filter(type == "Country/Area") %>%
  pull(region_subregion_country_area) %>%
  unique()

training_regions <- unique_countries %>%
  head(90)
predict_regions <- unique_countries %>%
  tail(47)

# Should be false:
any(training_regions %in% predict_regions)
```

```
## [1] FALSE
```

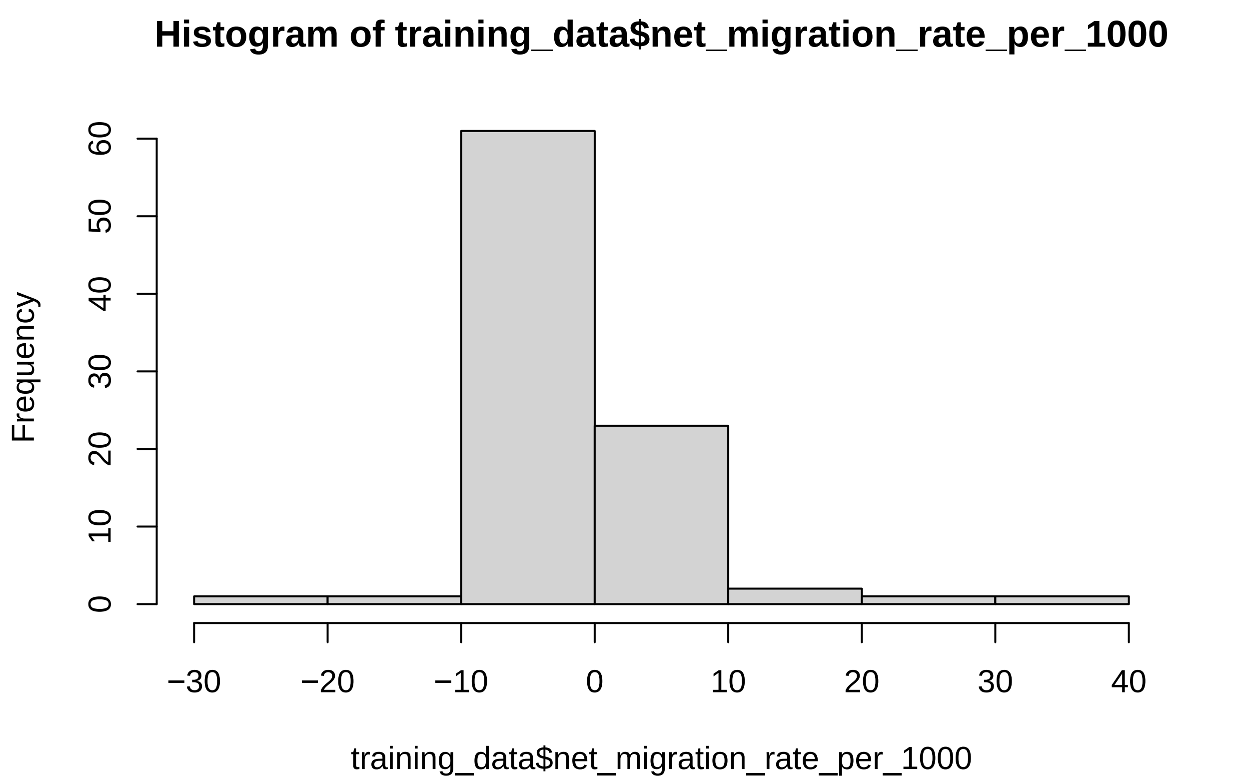
```r
training_data_2023 <- estimates %>%
  bind_rows(mediums) %>%
  filter(year == 2023) %>%
  filter(region_subregion_country_area %in% training_regions) %>%
  select(year, region_subregion_country_area, total_pop_january_thousands,
         crude_birth_rate_per_1000_pop, crude_death_rate_deaths_per_1000_population,
         net_migration_rate_per_1000)

training_data_2100 <- estimates %>%
  bind_rows(mediums) %>%
  filter(year == 2100) %>%
  filter(region_subregion_country_area %in% training_regions) %>%
  select(region_subregion_country_area, total_pop_january_thousands)

# Rename column for clarity
training_data_2100 <- training_data_2100 %>%
  rename(pop_2100 = total_pop_january_thousands)

training_data <- merge(x=training_data_2023,y=training_data_2100,
             by="region_subregion_country_area", all.x=TRUE)

# Look at histograms to see if any log transformations are needed
hist(training_data$total_pop_january_thousands)
```



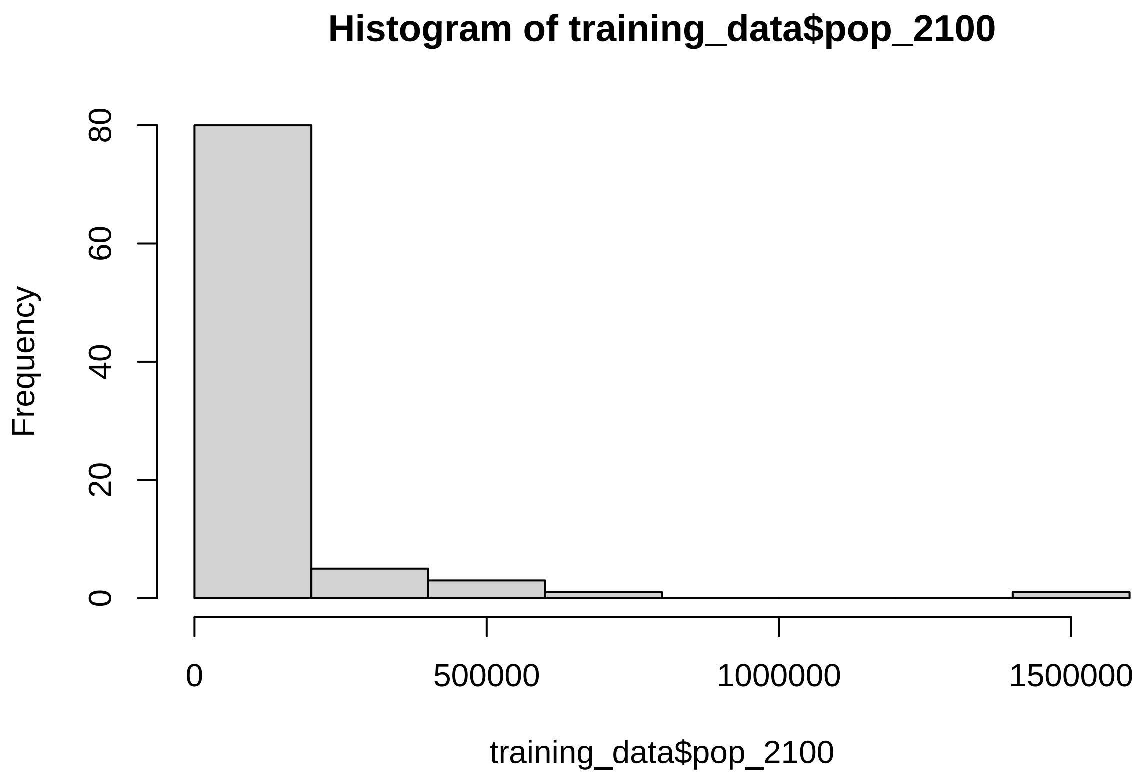

```r
hist(training_data$crude_birth_rate_per_1000_pop)
```


```r
hist(training_data$crude_death_rate_deaths_per_1000_population)
```


```r
hist(training_data$net_migration_rate_per_1000)
```




```r
hist(training_data$pop_2100)
```




```r
# total_pop_january_thousands and pop_2100 both have strong right skews
# so we will use log transformations for both

#linear model
lm_2100_pop <- lm(log(pop_2100) ~ log(total_pop_january_thousands) +
                                  crude_birth_rate_per_1000_pop +
                                  crude_death_rate_deaths_per_1000_population +
                                  net_migration_rate_per_1000,
                                  data = training_data)

summary(lm_2100_pop)
```

```
##
## Call:
## lm(formula = log(pop_2100) ~ log(total_pop_january_thousands) +
##     crude_birth_rate_per_1000_pop + crude_death_rate_deaths_per_1000_population +
##     net_migration_rate_per_1000, data = training_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.63431 -0.10207  0.00678  0.08552  0.34172
##
## Coefficients:
##                                               Estimate Std. Error t value
## (Intercept)                                  -0.543393   0.112802  -4.817
## log(total_pop_january_thousands)              0.994574   0.009198 108.131
## crude_birth_rate_per_1000_pop                 0.058937   0.001821  32.366
## crude_death_rate_deaths_per_1000_population -0.047161   0.009247  -5.100
## net_migration_rate_per_1000                   0.004756   0.002834   1.678
##                                              Pr(>|t|)
```

```
## (Intercept)                                  6.28e-06 ***
## log(total_pop_january_thousands)              < 2e-16 ***
## crude_birth_rate_per_1000_pop                 < 2e-16 ***
## crude_death_rate_deaths_per_1000_population 2.03e-06 ***
## net_migration_rate_per_1000                     0.097 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1754 on 85 degrees of freedom
## Multiple R-squared:  0.9939, Adjusted R-squared:  0.9937
## F-statistic:  3484 on 4 and 85 DF,  p-value: < 2.2e-16
```

```r
# Predicting future population of regions
predict_data_2023 <- estimates %>%
  bind_rows(mediums) %>%
  filter(year == 2023) %>%
  filter(region_subregion_country_area %in% predict_regions) %>%
  select(year, region_subregion_country_area, total_pop_january_thousands,
         crude_birth_rate_per_1000_pop, crude_death_rate_deaths_per_1000_population,
         net_migration_rate_per_1000)

predict_data_2100 <- estimates %>%
  bind_rows(mediums) %>%
  filter(year == 2100) %>%
  filter(region_subregion_country_area %in% predict_regions) %>%
  select(region_subregion_country_area, total_pop_january_thousands)

# Rename column for clarity
predict_data_2100 <- predict_data_2100 %>%
  rename(pop_2100 = total_pop_january_thousands)

predict_data <- merge(x=predict_data_2023,y=predict_data_2100,
             by="region_subregion_country_area", all.x=TRUE)

#Predicted values
predicted_values <- round(exp(predict(lm_2100_pop, newdata = predict_data)),3)

#pop_2100 is the UN's predicted population value for 2100, predicted_2100_pop is ours
# Put values of linear model into predicted_2100_pop
predict_data <- mutate(predict_data, predicted_2100_pop = predicted_values)
```

```r
mediums2100 <- mediums |> filter(year == 2100)
mediums2100$undata <- mediums2100$total_pop_january_thousands
predict_data$mlr_pop <- predict_data$total_pop_january_thousands
predict_data |> left_join(mediums2100, join_by(region_subregion_country_area))  |>
  select(region_subregion_country_area, mlr_pop, undata) |>
  mutate(diff = mlr_pop - undata,

         # mean differences between estimates and actual data:
         avg_diff = mean(diff),
         # median differences:
         med_diff = median(diff),
         sr = (undata - mlr_pop)^2,
         ssr = sum(sr),
         st = (undata - mean(undata))^2,
```

```r
        sst = sum(st),
        # calculating R^2 between UN model and predicted:
        r_sq = 1 - (ssr / sst)) |> select(-sr, -ssr, -st, -sst)
```

```
##         region_subregion_country_area    mlr_pop     undata       diff
## 1                      American Samoa     47.901     32.369     15.532
## 2                           Argentina  45459.024  38405.794   7053.230
## 3                           Australia  26320.802  43035.124 -16714.322
## 4                             Bermuda     64.724     35.827     28.897
## 5    Bolivia (Plurinational State of)  12159.495  17771.644  -5612.149
## 6                              Brazil 210707.000 163966.039  46740.961
## 7                              Canada  39059.725  53524.621 -14464.896
## 8                               Chile  19603.239  13507.688   6095.551
## 9                            Colombia  52032.604  47250.735   4781.869
## 10                       Cook Islands     14.477      7.918      6.559
## 11                            Ecuador  17902.009  19147.727  -1245.718
## 12        Falkland Islands (Malvinas)      3.483      2.254      1.229
## 13                               Fiji    921.747    881.621     40.126
## 14                      French Guiana    300.843    772.966   -472.123
## 15                   French Polynesia    280.771    187.920     92.851
## 16                          Greenland     55.962     37.351     18.611
## 17                               Guam    165.858    205.627    -39.769
## 18                          Guatemala  17984.483  25956.843  -7972.360
## 19                             Guyana    824.074    890.788    -66.714
## 20                           Honduras  10554.310  17039.536  -6485.226
## 21                           Kiribati    131.534    222.652    -91.118
## 22                   Marshall Islands     39.472     23.854     15.618
## 23                             Mexico 129171.119 130628.673  -1457.554
## 24        Micronesia (Fed. States of)    112.368    128.310    -15.942
## 25                              Nauru     11.845     20.645     -8.800
## 26                      New Caledonia    288.470    337.573    -49.103
## 27                        New Zealand   5151.357   5814.526   -663.169
## 28                          Nicaragua   6777.156   8630.243  -1853.087
## 29                               Niue      1.820      2.392     -0.572
## 30           Northern Mariana Islands     45.618     41.667      3.951
## 31                              Palau     17.748     11.160      6.588
## 32                             Panama   4430.030   5910.734  -1480.704
## 33                   Papua New Guinea  10296.209  18636.026  -8339.817
## 34                           Paraguay   6801.526   9058.714  -2257.188
## 35                               Peru  33656.344  38246.909  -4590.565
## 36          Saint Pierre and Miquelon      5.709      2.032      3.677
## 37                              Samoa    215.984    382.969   -166.985
## 38                    Solomon Islands    790.472   1847.317  -1056.845
## 39                           Suriname    626.028    710.256    -84.228
## 40                            Tokelau      2.342      4.373     -2.031
## 41                              Tonga    104.816    117.878    -13.062
## 42                             Tuvalu      9.904     12.256     -2.352
## 43           United States of America 342475.098 421007.222 -78532.124
## 44                            Uruguay   3388.682   2257.335   1131.347
## 45                            Vanuatu    316.732    873.271   -556.539
## 46 Venezuela (Bolivarian Republic of)  28250.783  28353.770   -102.987
## 47          Wallis and Futuna Islands     11.421      6.915      4.506
##     avg_diff med_diff     r_sq
## 1  -1879.935  -39.769 0.955747
```

```
## 2  -1879.935  -39.769 0.955747
## 3  -1879.935  -39.769 0.955747
## 4  -1879.935  -39.769 0.955747
## 5  -1879.935  -39.769 0.955747
## 6  -1879.935  -39.769 0.955747
## 7  -1879.935  -39.769 0.955747
## 8  -1879.935  -39.769 0.955747
## 9  -1879.935  -39.769 0.955747
## 10 -1879.935  -39.769 0.955747
## 11 -1879.935  -39.769 0.955747
## 12 -1879.935  -39.769 0.955747
## 13 -1879.935  -39.769 0.955747
## 14 -1879.935  -39.769 0.955747
## 15 -1879.935  -39.769 0.955747
## 16 -1879.935  -39.769 0.955747
## 17 -1879.935  -39.769 0.955747
## 18 -1879.935  -39.769 0.955747
## 19 -1879.935  -39.769 0.955747
## 20 -1879.935  -39.769 0.955747
## 21 -1879.935  -39.769 0.955747
## 22 -1879.935  -39.769 0.955747
## 23 -1879.935  -39.769 0.955747
## 24 -1879.935  -39.769 0.955747
## 25 -1879.935  -39.769 0.955747
## 26 -1879.935  -39.769 0.955747
## 27 -1879.935  -39.769 0.955747
## 28 -1879.935  -39.769 0.955747
## 29 -1879.935  -39.769 0.955747
## 30 -1879.935  -39.769 0.955747
## 31 -1879.935  -39.769 0.955747
## 32 -1879.935  -39.769 0.955747
## 33 -1879.935  -39.769 0.955747
## 34 -1879.935  -39.769 0.955747
## 35 -1879.935  -39.769 0.955747
## 36 -1879.935  -39.769 0.955747
## 37 -1879.935  -39.769 0.955747
## 38 -1879.935  -39.769 0.955747
## 39 -1879.935  -39.769 0.955747
## 40 -1879.935  -39.769 0.955747
## 41 -1879.935  -39.769 0.955747
## 42 -1879.935  -39.769 0.955747
## 43 -1879.935  -39.769 0.955747
## 44 -1879.935  -39.769 0.955747
## 45 -1879.935  -39.769 0.955747
## 46 -1879.935  -39.769 0.955747
## 47 -1879.935  -39.769 0.955747
```

Analysis Our intentions with creating this model were to try to estimate the data as best as we could with the fewest amount of predictors possible. After looking into how the UN recommended predicting data, we decided to create a model that accounted for birth rates, death rates, and migratory populations. In order to ensure our model met the proper assumptions of a linear regression model, we created histograms to depict the actual data and correct, via log transformations, to become more normally distributed. The two variables with a strong right skew, total population and the predicted populations, received a log transformation as a result. We also considered multicollinearity, especially between the life expectancy data and births and death. To minimize error, we omitted the variable from the model as it did not explain additional variability in the explanatory variable. As a result, we ended up with a model with a very high R^2, meaning that we were able to explain 99.39% of the variation in the the total population estimates with our chosen predictors. However, it is likely that the UN did not rely solely on a multiple linear regression model to inform their estimates. We can deduce this from the fact that our model used the same general categories as theirs, but the values do not completely align. Their methodology states that their probabilistic scenarios included uncertainty in their predictors, which differs from ours in that our data was only informed by one data set, and theirs incorporated the predictions and aggregates of 2,000. These discrepancies in methodology, chosen predictors, and in data collection are likely the cause of the differences in our estimates. Lastly, we used the UN model as the base in a secondary R^2 analysis, and the R^2 was 0.9557. In other words, our model was able to explain 95.57% of the variation in the UN data, with the remaining 4% a representation of their additional research. Overall, our model is a good approximation of that of the UN one, and can hopefully be used for external application.

United Nations, Department of Economic and Social Affairs, Population Division (2024). World Population Prospects 2024: Methodology of the United Nations population estimates and projections. UN DESA/POP/2024/DC/NO. 10, July 2024 [Advance unedited version].

6. Requirement-5 (1 pt)

Conclusion

Your analysis should conclude with a summary of key findings. I'm especially interested in any novel insights you uncover that go beyond the article's original conclusions.

Conclusion: Using a combination of EDA, data visualization techniques, and various R programming skills learned in Datasci 306 this semester, our group was able to come together to derive new and meaningful insights that expand upon the initial analysis provided by the article. By generating the twelve additional diagrams (shown in requirement 4 above) we are able to provide a deeper, more comprehensive understanding of the data provided to us. Our notable findings are documented below. To begin, we took a close look at Australia and the U.S. and discovered several key findings that go beyond the original conclusions in the article. First, for every decade since 1950, infant mortality rates are higher in the U.S. than in Australia. Secondly, we discover that there has been a convergence in female/male life expectancy between the U.S. and Australia. Thirdly, we add to the last diagram shown in the article that displays the annual net migration rates by looking at Australia and the U.S. in specific. We find that net migration rates are extremely volatile in Australia. Another notable discovery in our analysis is that mortality rates in Australia vary significantly by age. Furthermore, mortatlilty rates for people under 60 years of age have dramatically declined since 1950. Another novel insight that we uncover is that Africa is expected to have significant population growth over the next 77 years. We also reveal that Low-and-Lower-Middle income level countries demonstrate the largest variability in terms of the number of births. Furthermore, we discover that over the past two decades infant mortality rates have ranged from 5 to 20% in China as opposed to under 7% for the U.S.. As documented in the article, life expectancy rates are increasing and returning to pre-pandemic levels. Despite this increase in life expectancy rates, we use data visualization techniques to find that, on average, people in all countries of varying income levels are expected to have fewer children. That is, the average number of births is slowly falling (or in some cases it remains fairly constant). This supports diagram 3 from the article that shows declines in fertility rates. Furthermore, we take a closer look at life expectancy rates, but in a different dimension, and find that the gap in female life expectancy rates at age 15 has become more narrow overtime between less and more developed regions. Lastly, in generating our machine learning (ML) model we concluded that current population, birth rates, migration rates, and death rates are very significant predictors for future population. We reached the conclusion that the model we generated serves as a good approximation of the UN model.

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.

Submission

- You will upload the zip file containing finals.Rmd file and its PDF as a deliverable to Canvas. If you created a shiny app for predictions, you will add those files also to your zip file.
- You will present your findings by creating a video of a maximum 15 minutes duration, explaining the code and the workings of your project; all team members should explain their part in the project to receive credit. You will share the URL of the video on Canvas for us to evaluate. An ideal way to create this video would be to start a Zoom meeting, start recording, and then every member share their screen and explain their contribution.

It is not necessary to prepare slides (if you do it doesn't hurt) for the presentation. You may speak by showing the diagrams and/or code from your Posit project. Every team member should explain their part in the project along with the insights they derived by explaining the charts and summaries for full credit to each member.

Your project will be evaluated for clean code, meaningful/insightful EDA and predictions.

Note:

- Each plot must be accompanied by a summary that clarifies the rationale behind its creation and what insights the plot unveils. Every diagram should possess standalone significance, revealing something more compelling than the other charts
- After the deadline, instructors will select the top three outstanding analytics projects. The teams responsible for these exceptional analyses will have their video shared with the class

We will not accept submissions after the deadline; December 10th 4 pm