DATASCI 306, Fall 2024, Final Group Project

Group 40 - Matthew Varela, Matthew Suba, Weikang Hu, Adrian Vergara, Siya Modi

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.

```
file_path <- "data/WPP2024_GEN_F01_DEMOGRAPHIC_INDICATORS_COMPACT.xlsx"

rough_estimates_df <- read_excel(file_path, sheet = "Estimates", skip = 16, col_types = "text")

rough_medium_variant_df <- read_excel(file_path, sheet = "Medium variant", skip = 16, col_types = "text
```

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

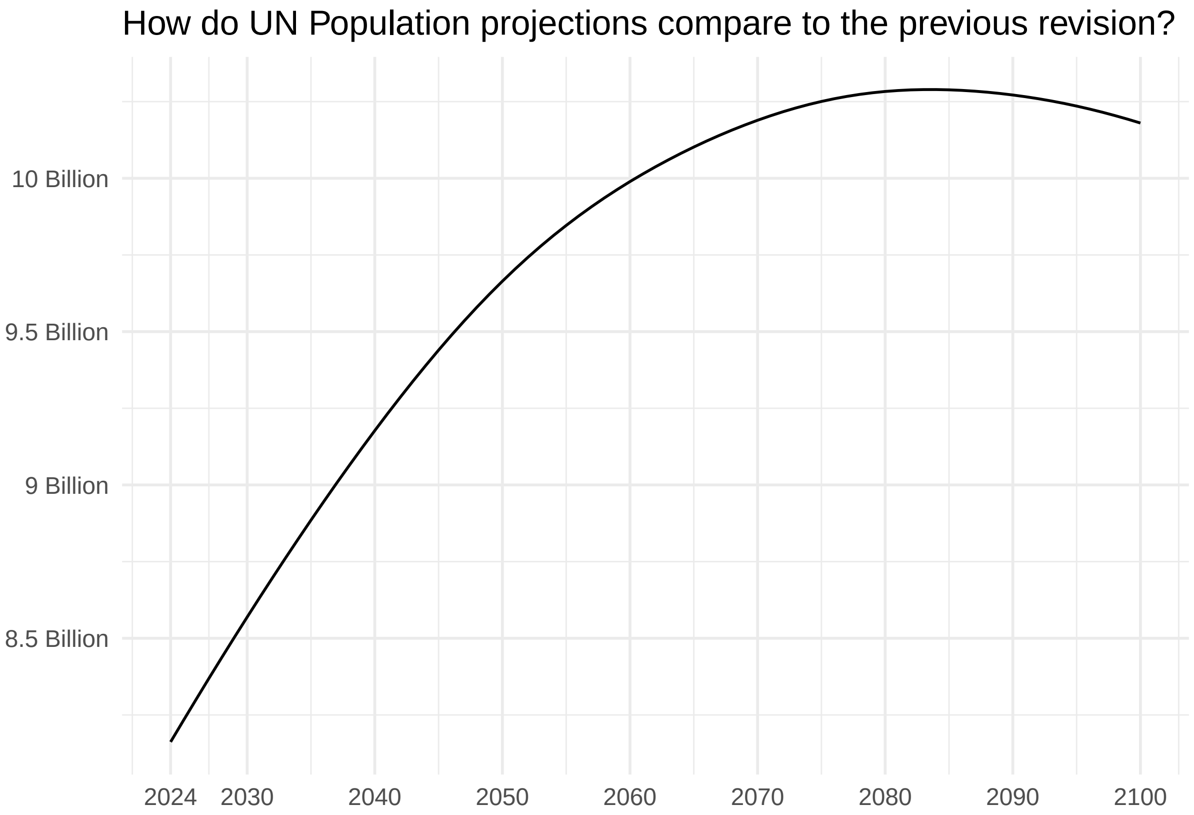
You could also remove the countries/regions that you are not interested in exploring in this step and re-save a smaller file in the same `data` folder, with a different name so that working with it becomes easier going forward.

Explain your reasoning for each clean up step.

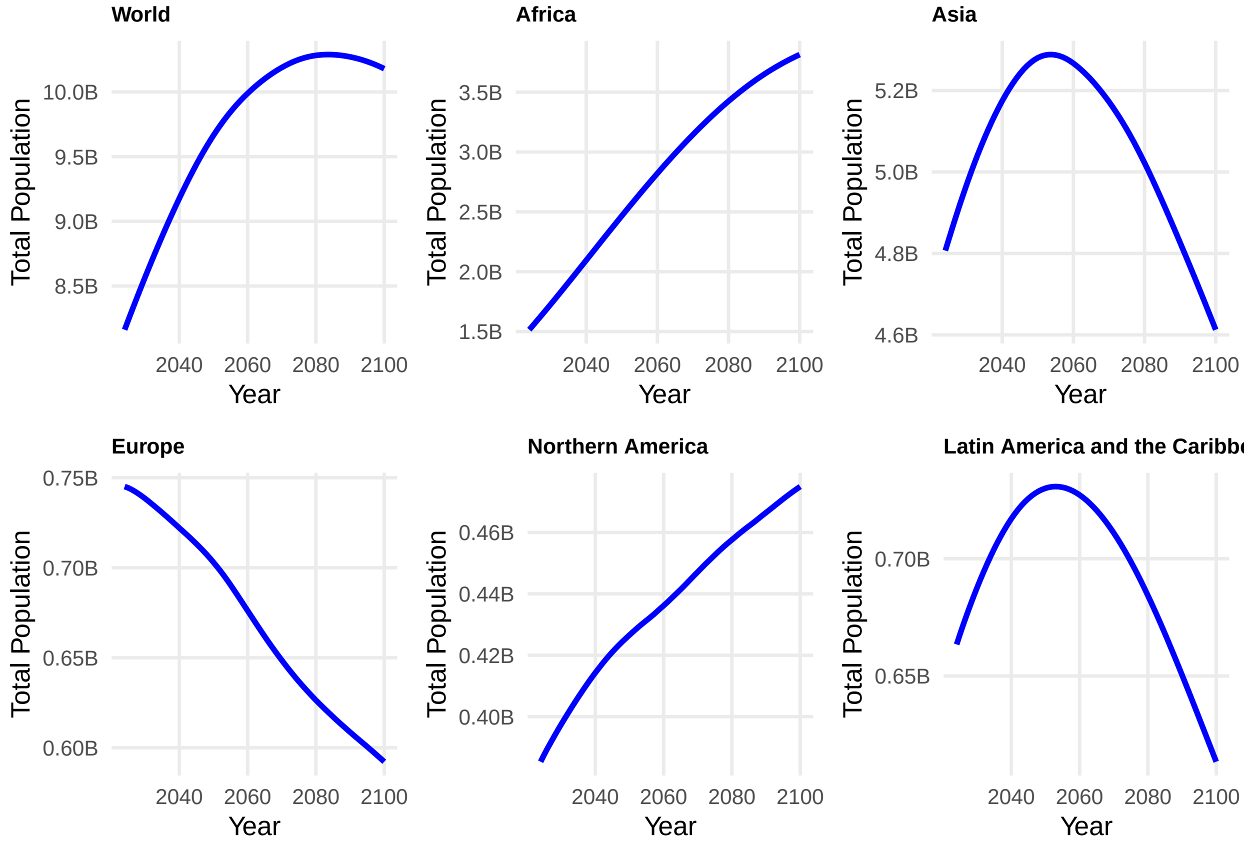
```r
column_rename <- function(df) {
  df |>
    rename_with(~ gsub(" ", "_", .)) |>
    rename_with(~ gsub("[^[:alnum:]_]", "", .)) |>
    rename_with(~ tolower(.))
}

estimates_df <- column_rename(rough_estimates_df)
medium_variant_df <- column_rename(rough_medium_variant_df)

estimates_df <- estimates_df %>%
  filter(type != 'Label/Separator')

estimates_df <- estimates_df %>%
  select(-index, -variant, -sdmx_code, -parent_code)

medium_variant_df <- medium_variant_df %>%
  filter(type != 'Label/Separator')

medium_variant_df <- medium_variant_df %>%
  select(-index, -variant, -sdmx_code, -parent_code)

estimates_df <- estimates_df %>%
  rename(
    region = region_subregion_country_or_area_,
    total_pop_jan = total_population_as_of_1_january_thousands,
    total_pop_july = total_population_as_of_1_july_thousands,
    male_pop_july = male_population_as_of_1_july_thousands,
    female_pop_july = female_population_as_of_1_july_thousands,
    pop_density = population_density_as_of_1_july_persons_per_square_km,
    pop_sex_ratio = population_sex_ratio_as_of_1_july_males_per_100_females,
    median_age = median_age_as_of_1_july_years,
    natural_change = natural_change_births_minus_deaths_thousands,
    rate_natural_change = rate_of_natural_change_per_1000_population,
    pop_change = population_change_thousands,
    pop_growth_rate = population_growth_rate_percentage,
    pop_doubling_time = population_annual_doubling_time_years,
    births = births_thousands,
    births_women_15_19 = births_by_women_aged_15_to_19_thousands,
    crude_birth_rate = crude_birth_rate_births_per_1000_population,
    total_fertility_rate = total_fertility_rate_live_births_per_woman,
    net_reproduction_rate = net_reproduction_rate_surviving_daughters_per_woman,
    mean_age_childbearing = mean_age_childbearing_years,
    sex_ratio_at_birth = sex_ratio_at_birth_males_per_100_female_births,
    total_deaths = total_deaths_thousands,
    male_deaths = male_deaths_thousands,
    female_deaths = female_deaths_thousands,
    crude_death_rate = crude_death_rate_deaths_per_1000_population,
    life_exp_birth_both = life_expectancy_at_birth_both_sexes_years,
    life_exp_birth_male = male_life_expectancy_at_birth_years,
    life_exp_birth_female = female_life_expectancy_at_birth_years,
    net_migrants = net_number_of_migrants_thousands,
    net_migration_rate = net_migration_rate_per_1000_population
  )

medium_variant_df <- medium_variant_df %>%
  rename(
    region = region_subregion_country_or_area_,
    total_pop_jan = total_population_as_of_1_january_thousands,
    total_pop_july = total_population_as_of_1_july_thousands,
    male_pop_july = male_population_as_of_1_july_thousands,
    female_pop_july = female_population_as_of_1_july_thousands,
    pop_density = population_density_as_of_1_july_persons_per_square_km,
    pop_sex_ratio = population_sex_ratio_as_of_1_july_males_per_100_females,
    median_age = median_age_as_of_1_july_years,
    natural_change = natural_change_births_minus_deaths_thousands,
    rate_natural_change = rate_of_natural_change_per_1000_population,
    pop_change = population_change_thousands,
    pop_growth_rate = population_growth_rate_percentage,
    pop_doubling_time = population_annual_doubling_time_years,
    births = births_thousands,
    births_women_15_19 = births_by_women_aged_15_to_19_thousands,
    crude_birth_rate = crude_birth_rate_births_per_1000_population,
    total_fertility_rate = total_fertility_rate_live_births_per_woman,
    net_reproduction_rate = net_reproduction_rate_surviving_daughters_per_woman,
    mean_age_childbearing = mean_age_childbearing_years,
    sex_ratio_at_birth = sex_ratio_at_birth_males_per_100_female_births,
    total_deaths = total_deaths_thousands,
    male_deaths = male_deaths_thousands,
    female_deaths = female_deaths_thousands,
    crude_death_rate = crude_death_rate_deaths_per_1000_population,
    life_exp_birth_both = life_expectancy_at_birth_both_sexes_years,
    life_exp_birth_male = male_life_expectancy_at_birth_years,
    life_exp_birth_female = female_life_expectancy_at_birth_years,
    net_migrants = net_number_of_migrants_thousands,
    net_migration_rate = net_migration_rate_per_1000_population
  )



numeric_cols <- c(
  'year',
  'total_pop_jan',
  'total_pop_july',
  'male_pop_july',
  'female_pop_july',
  'pop_density',
  'pop_sex_ratio',
  'median_age',
  'natural_change',
  'rate_natural_change',
  'pop_change',
  'pop_growth_rate',
  'pop_doubling_time',
  'births',
  'births_women_15_19',
  'crude_birth_rate',
  'total_fertility_rate',
  'net_reproduction_rate',
  'mean_age_childbearing',
  'sex_ratio_at_birth',
  'total_deaths',
  'male_deaths',
  'female_deaths',
  'crude_death_rate',
  'life_exp_birth_both',
  'life_exp_birth_male',
  'life_exp_birth_female',
  'net_migrants',
  'net_migration_rate'
)

non_numeric_patterns <- c("^\\.{2,}$", "^\\-$", "^\\s*$", "^-$")

estimates_df[numeric_cols] <- estimates_df[numeric_cols] %>%
  mutate(across(everything(), ~ ifelse(str_detect(., paste(non_numeric_patterns, collapse = "|")), NA,

estimates_df <- estimates_df %>%
  mutate(across(all_of(numeric_cols), as.numeric))

medium_variant_df[numeric_cols] <- medium_variant_df[numeric_cols] %>%
  mutate(across(everything(), ~ ifelse(str_detect(., paste(non_numeric_patterns, collapse = "|")), NA,

medium_variant_df <- medium_variant_df %>%
  mutate(across(all_of(numeric_cols), as.numeric))
```

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.

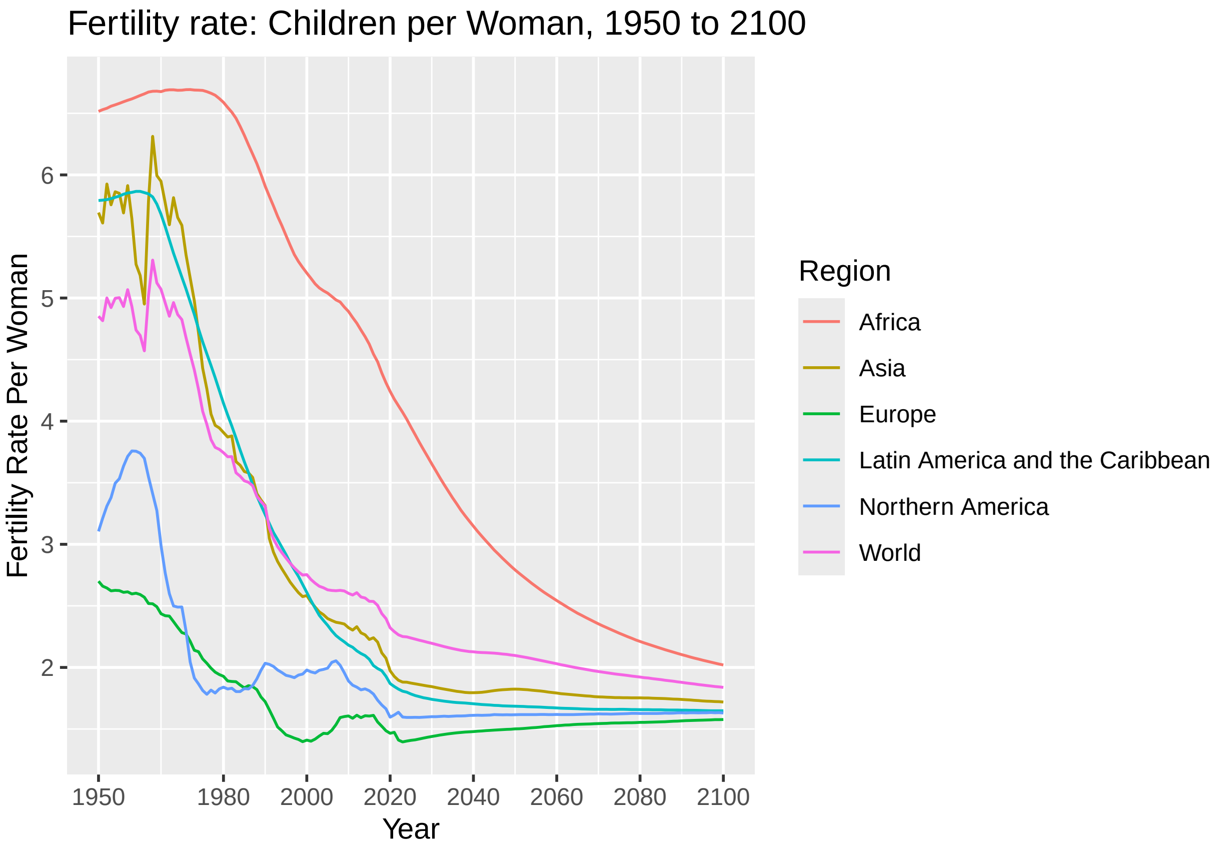
```
medium_variant_df |>
  filter(`region` == "World") |>
  ggplot(aes(x = year, y = `total_pop_july` /1e6 )) +
  geom_line() +
  scale_y_continuous(
    breaks = c(8.5, 9, 9.5, 10),
    labels = c("8.5 Billion", "9 Billion", "9.5 Billion", "10 Billion")
  ) +
  scale_x_continuous(
    limits = c(2024, 2100),
    breaks = c(2024, 2030, 2040, 2050, 2060, 2070, 2080, 2090, 2100)
  ) +
  theme_minimal()+
  labs(title = "How do UN Population projections compare to the previous revision?",
       x = "", y = "")
```



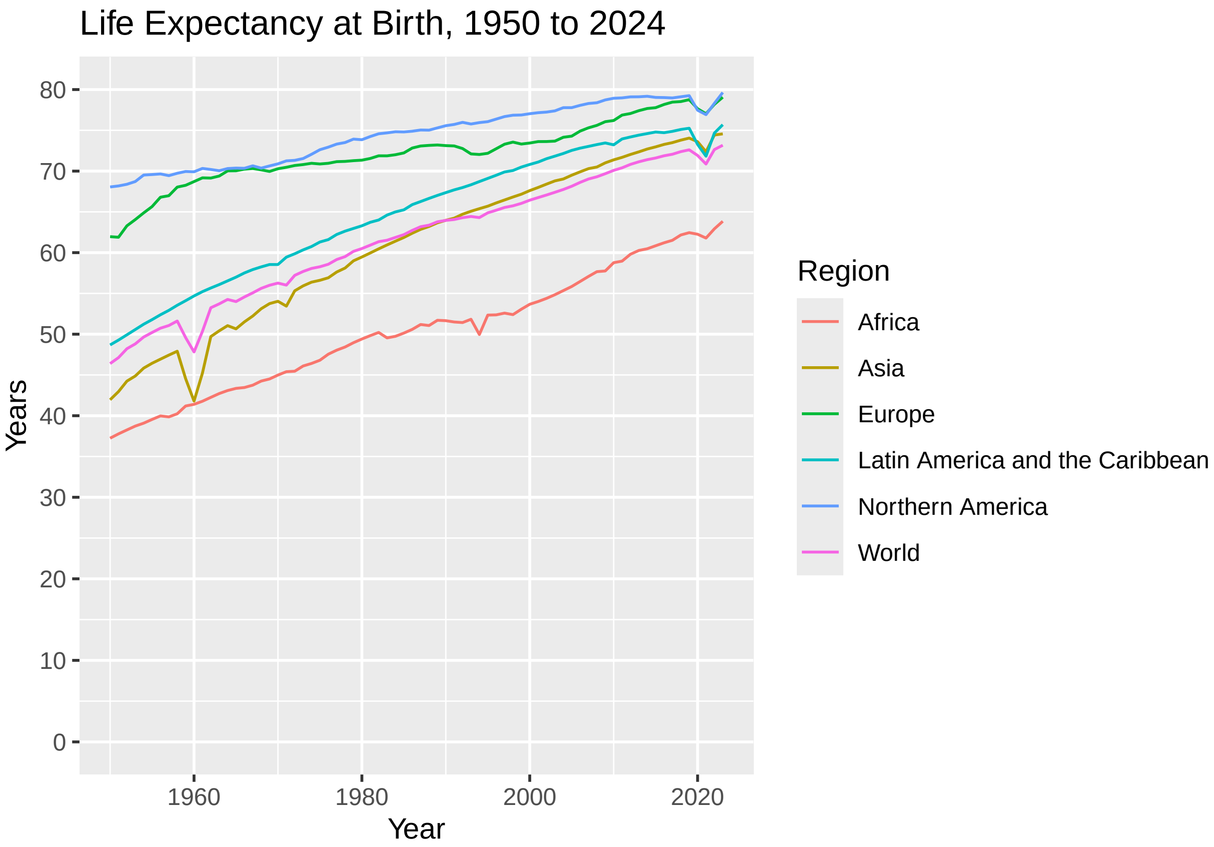
```
regions <- c('World', 'Africa', 'Asia', 'Europe', 'Northern America', 'Latin America and the Caribbean'

plot_list <- list()

for (i in seq_along(regions)) {
  region_name <- regions[i]

  region_data <- medium_variant_df %>%
    filter(region == region_name)

  if (nrow(region_data) == 0) {
    cat("No data available for", region_name, "\n")
    next
  }
  p <- ggplot(region_data, aes(x = year, y = total_pop_july)) +
    geom_line(color = "blue", linewidth = 1) +
    labs(
      title = region_name,
      x = "Year",
      y = "Total Population"
    ) +
    scale_y_continuous(labels = scales::number_format(scale = 1e-6, suffix = "B")) +
    scale_x_continuous(limits = c(2024, 2100)) +
    theme_minimal() +
    theme(
      plot.title = element_text(size = 8, face = "bold"),
      axis.title = element_text(size = 10),
      axis.text = element_text(size = 8),
      panel.grid.major = element_line(linewidth = 0.6),
      panel.grid.minor = element_blank()
    )
  plot_list[[i]] <- p
}
do.call(grid.arrange, c(plot_list, ncol = 3))
```



```
regions <- c('World', 'Africa', 'Asia', 'Europe', 'Northern America', 'Latin America and the Caribbean'

combined <- rbind(estimates_df, medium_variant_df)

combined |> filter(region %in% c("World", "Europe", "Africa", "Asia", "Northern America", "Latin Americ
  ggplot(aes(x = year, y = total_fertility_rate,
             color = region)) +
  geom_line() +
  labs(title = "Fertility rate: Children per Woman, 1950 to 2100", x = "Year", y = "Fertility Rate Per 
  scale_y_continuous(breaks = 1:6) +
  scale_x_continuous(
    breaks = c(1950, seq(1980, 2100, 20)))
```



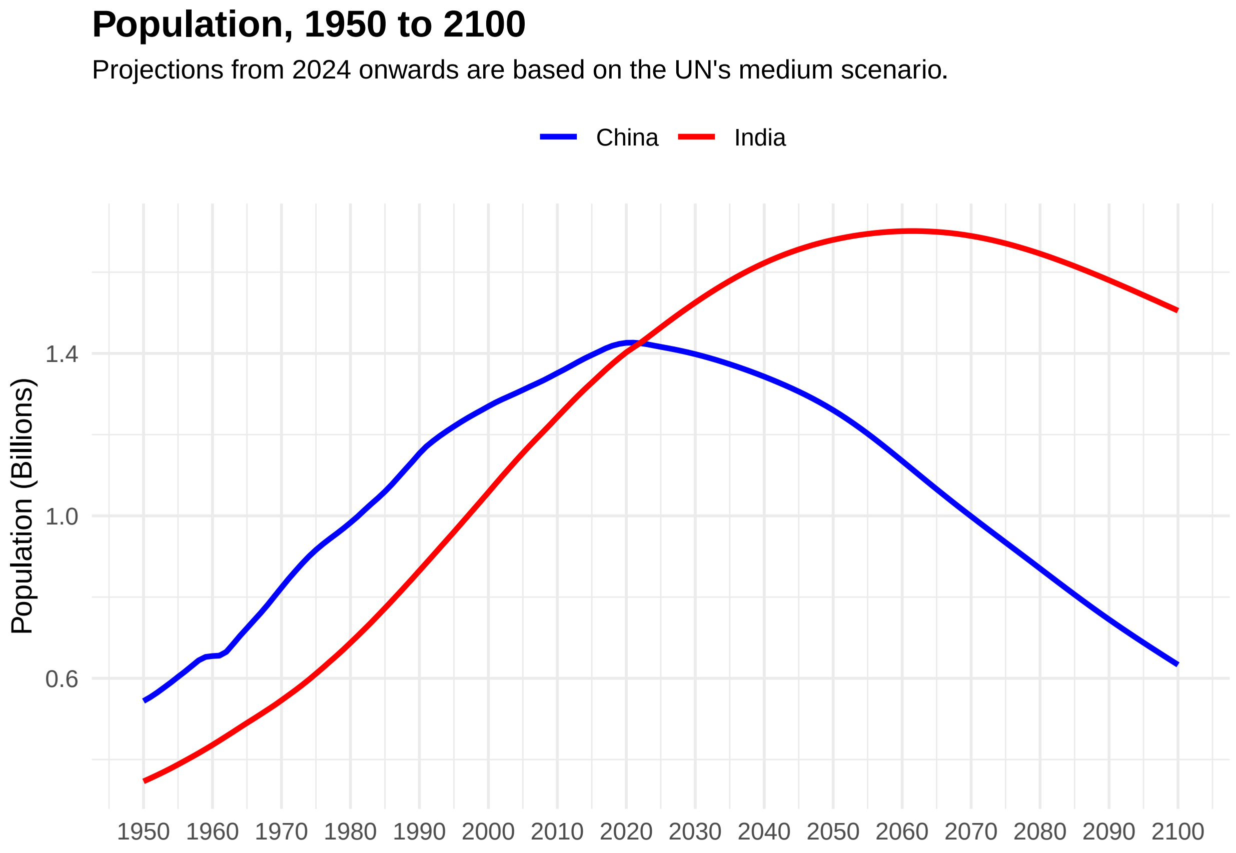
```
estimates_df |> filter(region %in% c("World", "Latin America and the Caribbean", "Europe", "Africa", "No
  geom_line() +
  labs(title = "Life Expectancy at Birth, 1950 to 2024", x = "Year", y = "Years", color = "Region") +
  scale_y_continuous(breaks = seq(0, 100, by = 10), limits = c(0, 80))
```



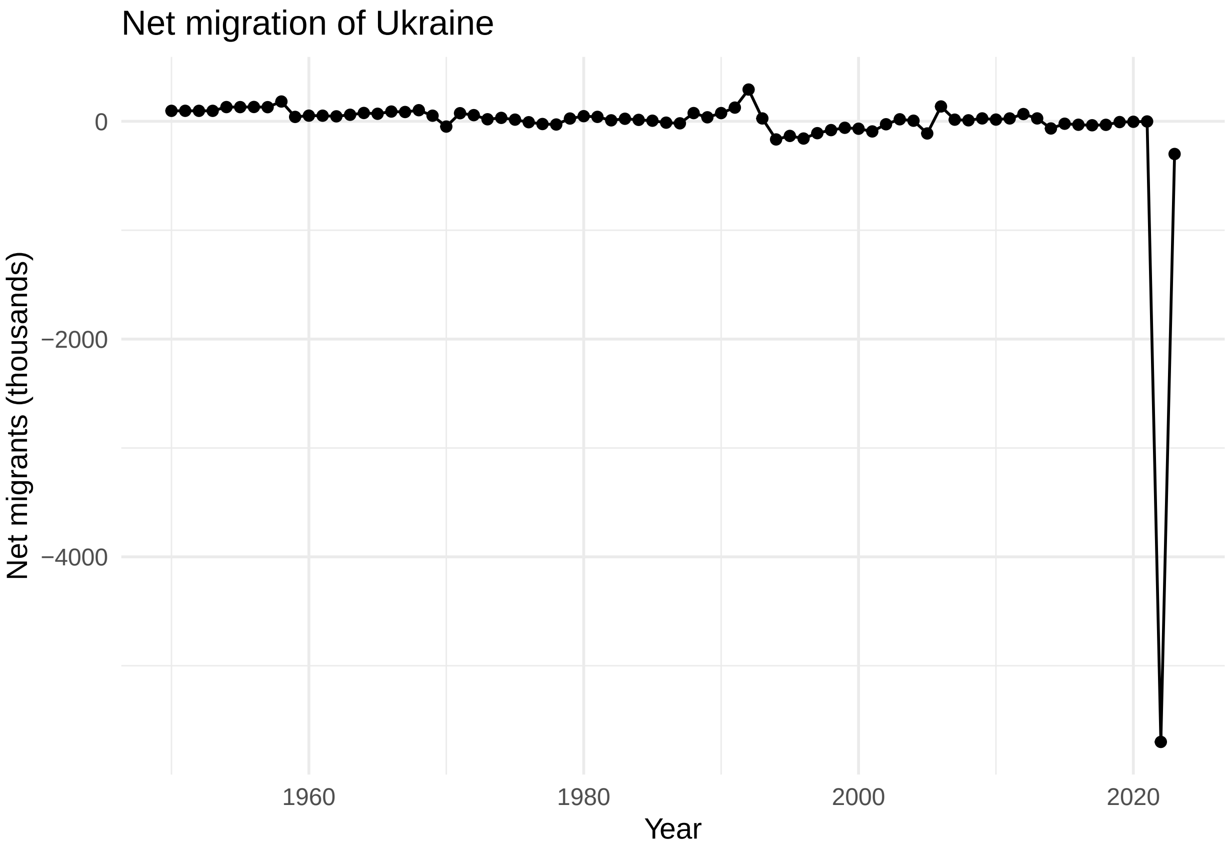
```
filtered_data1 <- estimates_df %>% filter(region%in% c("India","China")) %>%
  select(region, year, total_pop_july)
filtered_data1 <- filtered_data1 %>%
  mutate(total_pop_july = total_pop_july / 1e6)

filtered_data <- medium_variant_df %>%
  filter(region %in% c("India", "China")) %>%
  select(region, year, total_pop_july)

filtered_data <- filtered_data %>%
  mutate(total_pop_july = total_pop_july / 1e6)

combined_data <- bind_rows(filtered_data,filtered_data1)
ggplot(combined_data, aes(x = year, y = total_pop_july, color = region)) +
  geom_line(size = 1) +
  scale_color_manual(values = c("India" = "red", "China" = "blue")) +
  scale_x_continuous(
    limits = c(1950, 2100),
    breaks = seq(1950, 2100, by = 10)
  ) +
  labs(
    title = "Population, 1950 to 2100",
    subtitle = "Projections from 2024 onwards are based on the UN's medium scenario.",
    x = NULL,
    y = "Population (Billions)",
    color = NULL
  ) +
```

```r
  theme_minimal() +
  theme(
    plot.title = element_text(size = 14, face = "bold"),
    plot.subtitle = element_text(size = 10),
    legend.position = "top"
  )
```

```
## 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.
```



```r
estimates_df |>
  filter(region == "Ukraine") |>
  select(year, net_migrants) |>
  ggplot(aes(x=year, y=net_migrants)) +
  geom_line() +
  theme_minimal() +
  labs(title = "Net migration of Ukraine",
       x = "Year",
       y = "Net migrants (thousands)") +
  geom_point()
```

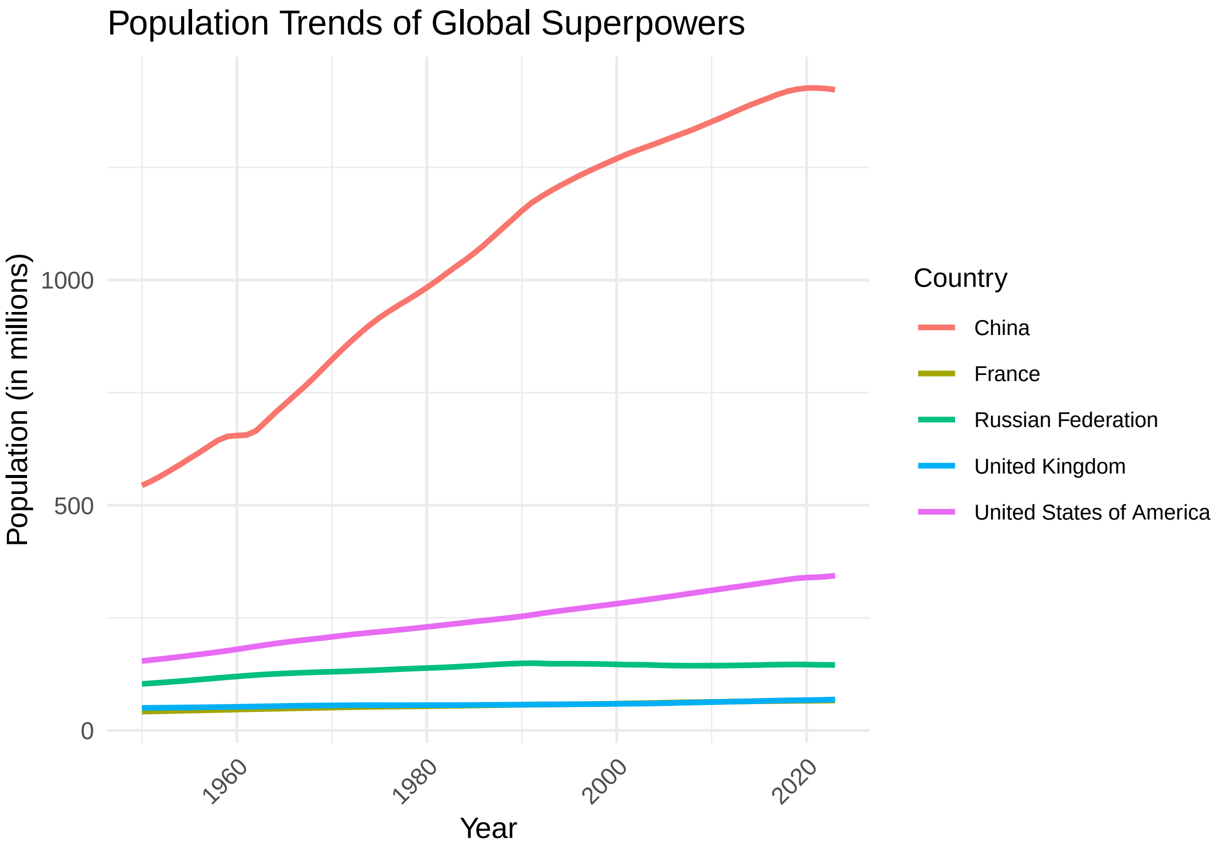


4. **Requirement-4 (12 pt)**

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.

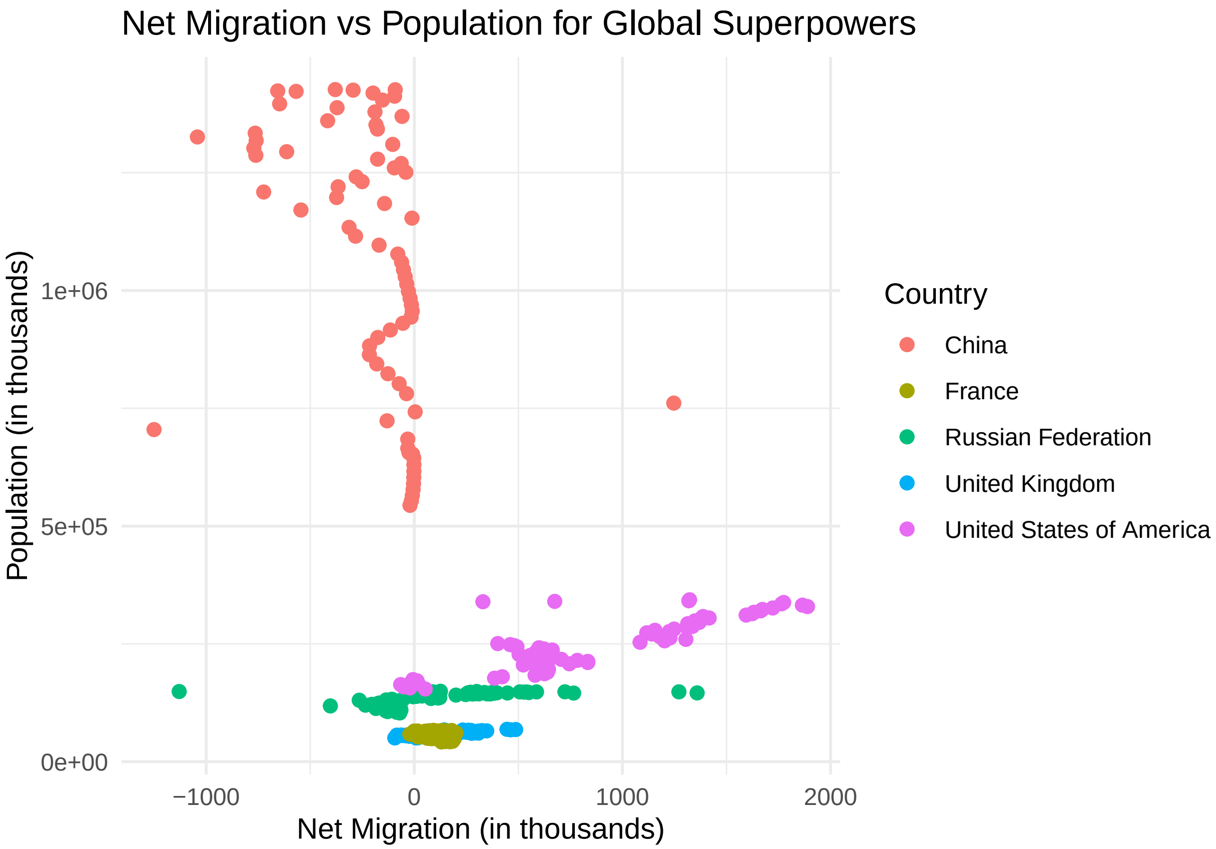
Summarize your interpretations after each chart.

```
countries <- c("China", "France", "Russian Federation", "United Kingdom", "United States of America")

global_superpowers_data <- estimates_df %>%
  filter(region %in% countries)

global_superpowers_data <- global_superpowers_data %>%
  mutate(
    year = as.numeric(year),
    total_pop_july = as.numeric(total_pop_july)
  )

ggplot(global_superpowers_data, aes(x = year, y = total_pop_july / 1e3, color = region)) +
  geom_line(size = 1.0) +
  labs(
    title = "Population Trends of Global Superpowers",
    x = "Year",
    y = "Population (in millions)",
    color = "Country"
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1),
    legend.title = element_text(size = 10),
    legend.text = element_text(size = 8)
  )
```



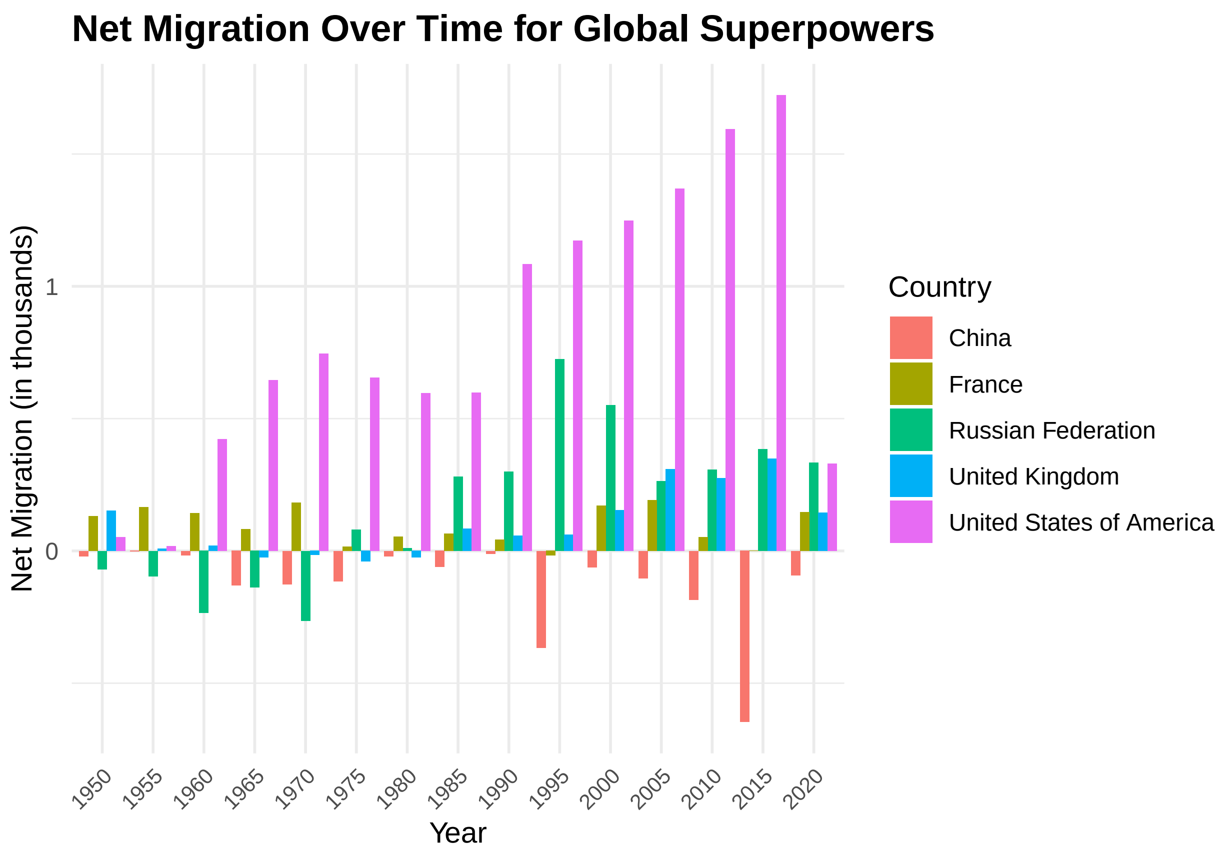
Explanation 1 From this, we can see and compare the overall populations between the three counties over time. It is clear that the population in China is much larger than that of the France, Russia, the UN or the United States. In addition, we can see how the Chinese population increases faster over time than the other four countries indicating a greater growth rate. China and the United States seem to have a positive growth rate as they are both increasing over time, and Russia, the United Kingdom and and France appear flat out over time with minimal growth.

```
ggplot(global_superpowers_data, aes(x = net_migrants, y = total_pop_july, color = region)) +
  geom_point(size = 2) +
  labs(
    title = "Net Migration vs Population for Global Superpowers",
    x = "Net Migration (in thousands)",
    y = "Population (in thousands)",
    color = "Country"
  ) +
  theme_minimal()
```



Explanation 2 This scatter plot illustrates the relationship between population and net migration of the global super power countries and how much they influence each other. Since dots for China show a vertical linear relationship, it is evident net migration has little to no influence over the overall population, as it remains close to or below 0, and overall Chinese population had grown, as seen in the graph above. Meanwhile, the United States shows the strongest positive linear relationship. This means that an increase in net migration is more likely to cause an increase in overall population. The other three show migration having little to no influence over population.

```
migration_data <- global_superpowers_data %>%
  filter(!is.na(net_migrants))

migration_data <- migration_data %>%
  filter(year %% 5 == 0)

ggplot(migration_data, aes(x = factor(year), y = net_migrants / 1e3, fill = region)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(
    title = "Net Migration Over Time for Global Superpowers",
    x = "Year",
    y = "Net Migration (in thousands)",
    fill = "Country"
  ) +
  scale_y_continuous(labels = comma) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1, size = 8),
    plot.title = element_text(size = 14, face = "bold")
  )
```



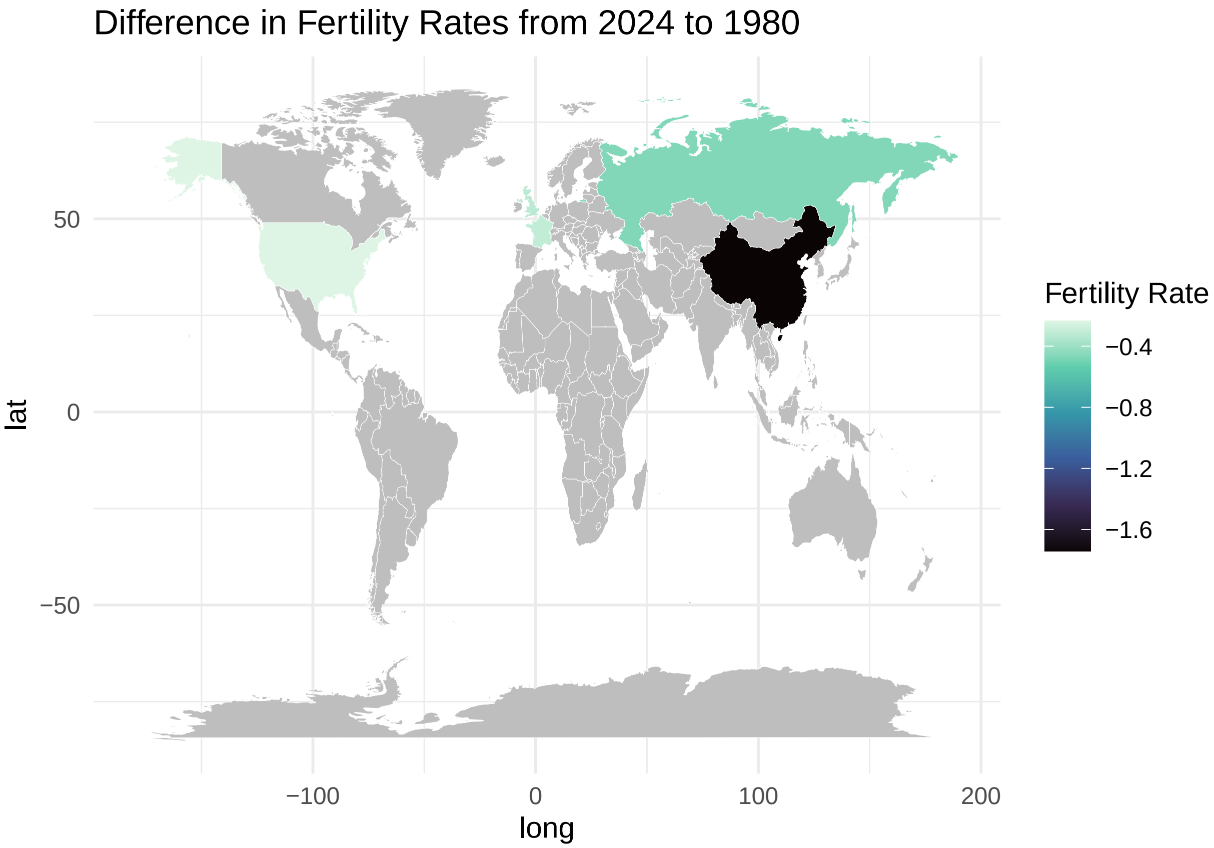
Explanation 3 In this chart we can see the net migration for the countries over time. It is clear that the United states has the greatest number of migrants into the country for the majority of the years, while China experiences the most migration out of the country since their number is almost always negative. It is interesting to see how these patterns differ from overall population with China dominating greatly as number of migrants seems to have minimal effect in the total population.

```
world_map = map_data("world")

world_map <- world_map |>
  mutate(iso3_code = countrycode(sourcevar = region, origin = "country.name", destination = "iso3c"))
```

```
## Warning: There was 1 warning in `mutate()`.
## i In argument: `iso3_code = countrycode(sourcevar = region, origin =
##   "country.name", destination = "iso3c")`.
## Caused by warning:
## ! Some values were not matched unambiguously: Ascension Island, Azores, Barbuda, Bonaire, Canary Isla
```

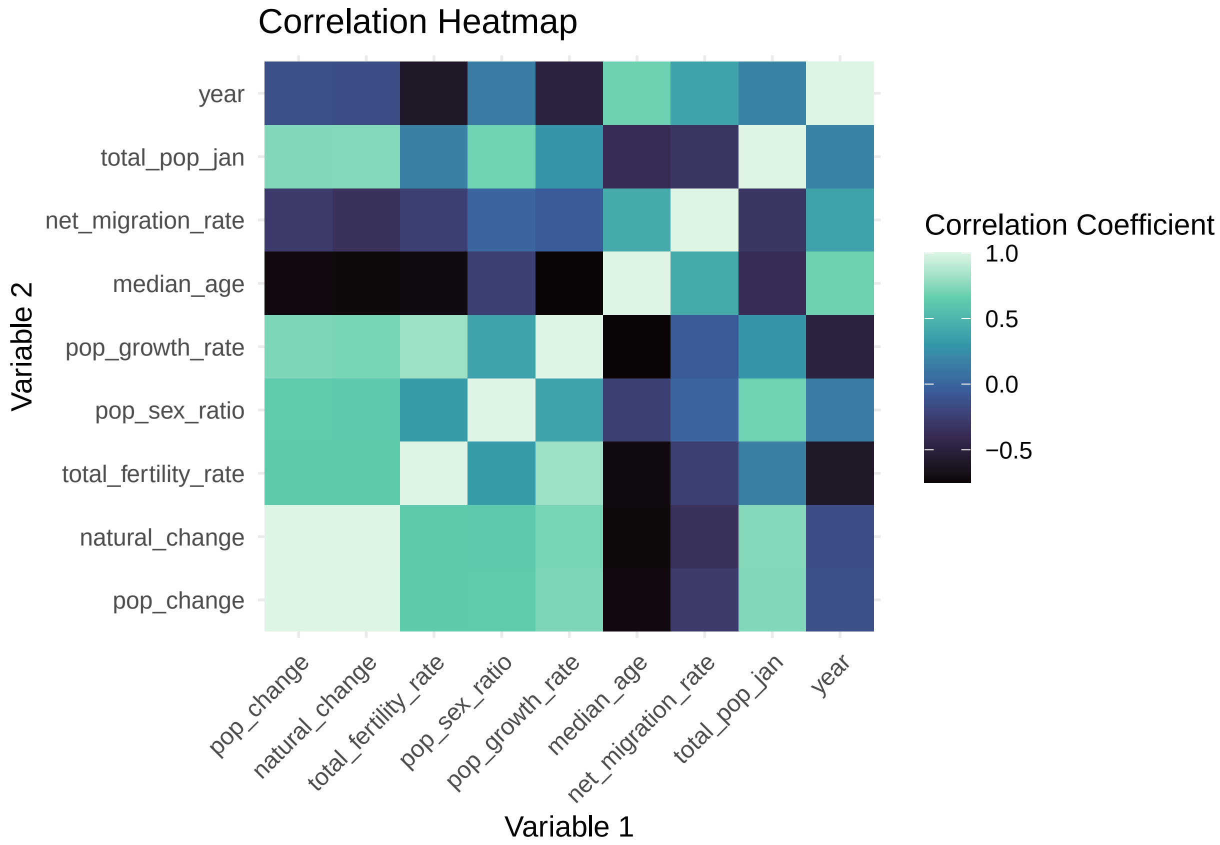
```
fertility_data <- global_superpowers_data |>
  filter(!is.na(total_fertility_rate)) |>
  mutate(iso3_code = countrycode(sourcevar = region, origin = "country.name", destination = "iso3c"))

fertility_data <- fertility_data |>
  filter(year %in% c(1980, 2023)) |>
  select(region, iso3_code, total_fertility_rate, year)|>
  pivot_wider(names_from = year, values_from = total_fertility_rate, names_prefix = "year_") |>
  mutate(fertility_diff =  year_2023 - year_1980)

world_map <- world_map |>
  left_join(fertility_data, by = "iso3_code")
```

```
ggplot(data = world_map, aes(x = long, y = lat, group = group, fill = fertility_diff)) +
  geom_polygon(color = "white", size = 0.1) +
  scale_fill_viridis_c(option = "mako", na.value = "grey", name = "Fertility Rate") +
  theme_minimal() +
  labs(
    title = "Difference in Fertility Rates from 2024 to 1980"
  )
```



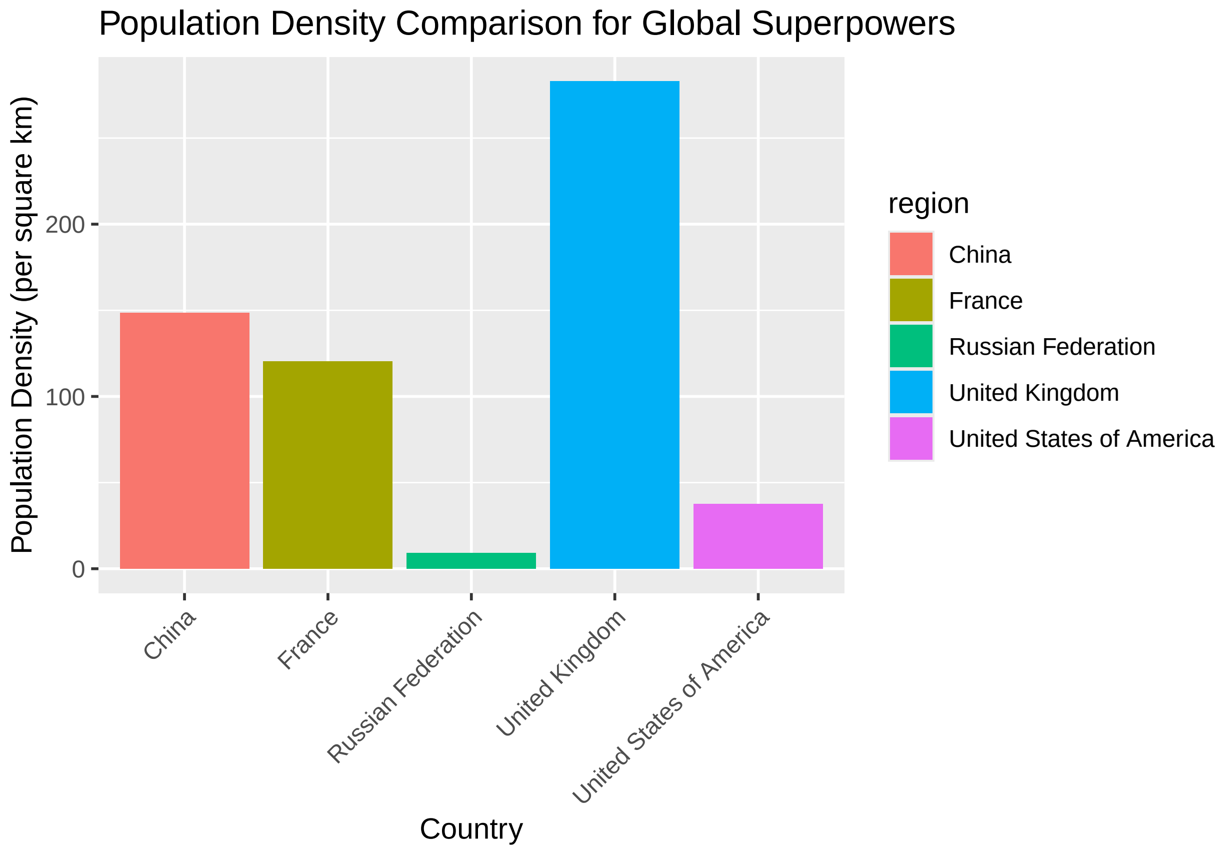
Explanation 4 This map visualized the changing fertility rates from 1980 to 2024. All of the countries we focused on have declining fertility rates over time, some much more than others. This makes sense as family size has generally decreased over time, some possible factors could include increased urbanization. It is also shown how China has the highest declining rate compared to the rest of the countries involved, which could highlight strict policies such as the one-child policy, which was implemented to address their rapid population growth, that we can see in graph 1. The remainder of the countries seem to experience decreasing fertility rates, but with much less severity.

```
filtered_data <- global_superpowers_data |>
  filter(region %in% countries) |>
  select(pop_change, natural_change, total_fertility_rate, pop_sex_ratio,
         pop_growth_rate, median_age, net_migration_rate, total_pop_jan, year)

correlation_matrix <- cor(filtered_data) |>
  melt(varnames = c("Var1", "Var2"), value.name = "Correlation")

ggplot(correlation_matrix, aes(x = Var1, y = Var2, fill = Correlation)) +
  geom_tile() +
  scale_fill_viridis_c(option = "mako") +
  theme_minimal() +
  labs(title = "Correlation Heatmap",
       x = "Variable 1",
       y = "Variable 2",
       fill = "Correlation Coefficient") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



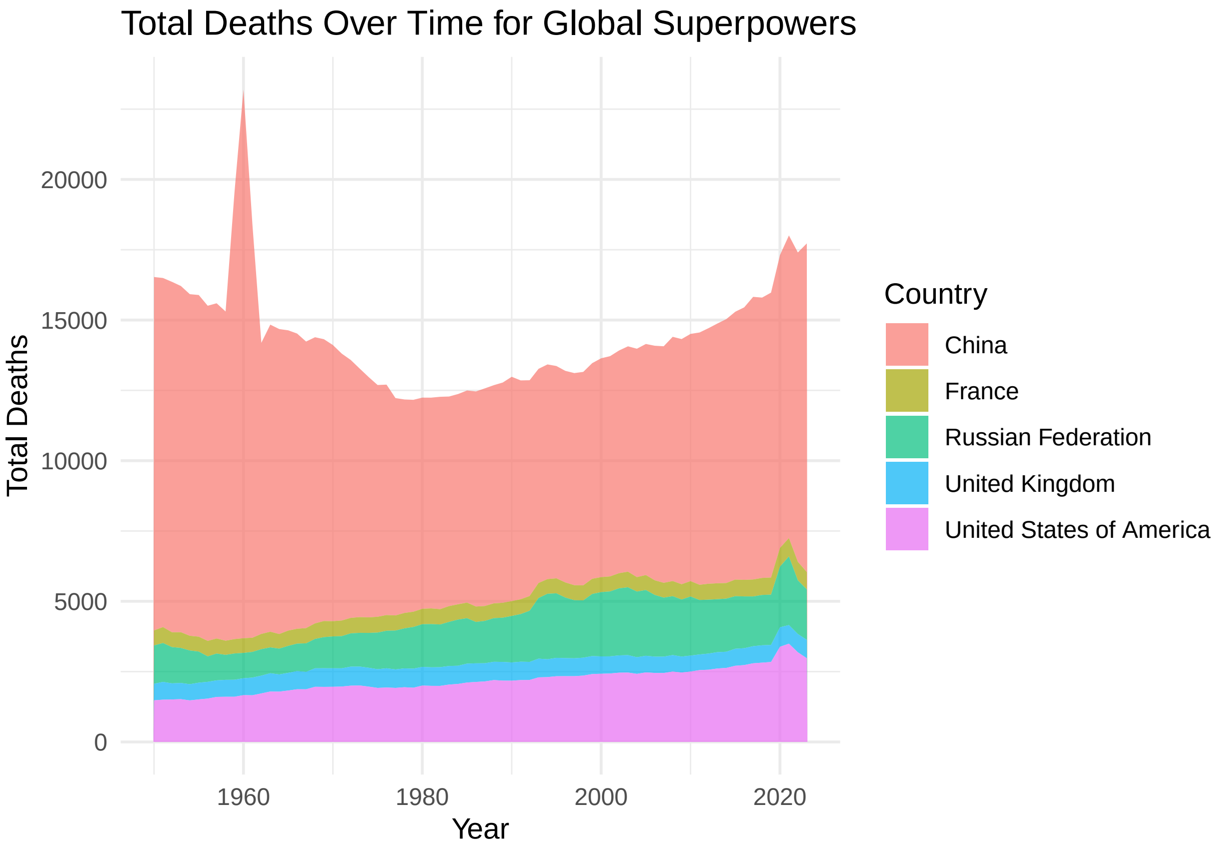
Explanation 5 This correlation heat map provides an effective overview into various demographic variables, and highlighting the strength of correlation between them. The lighter colors, or relationships with correlation coefficients close to 1 tend to increase together and have a strong linear relationship. One example of this is total fertility rate and population growth rate. An increase in total fertility rate has great influence in population causing an increase in population growth rate. Meanwhile variables that are colored darker purple have inverse correlations, meaning as one variable increases, the other tends to decrease. Lastly, variables such as year and population change are blue with a correlation coefficient close to zero. This means that these variables have very little influence over each other and imply little to no linear relationship.

```
ggplot(global_superpowers_data, aes(x = region, y = pop_density, fill = region)) +
geom_bar(stat = "identity", position = "dodge") +
labs(title = "Population Density Comparison for Global Superpowers",
x = "Country", y = "Population Density (per square km)") +
   theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



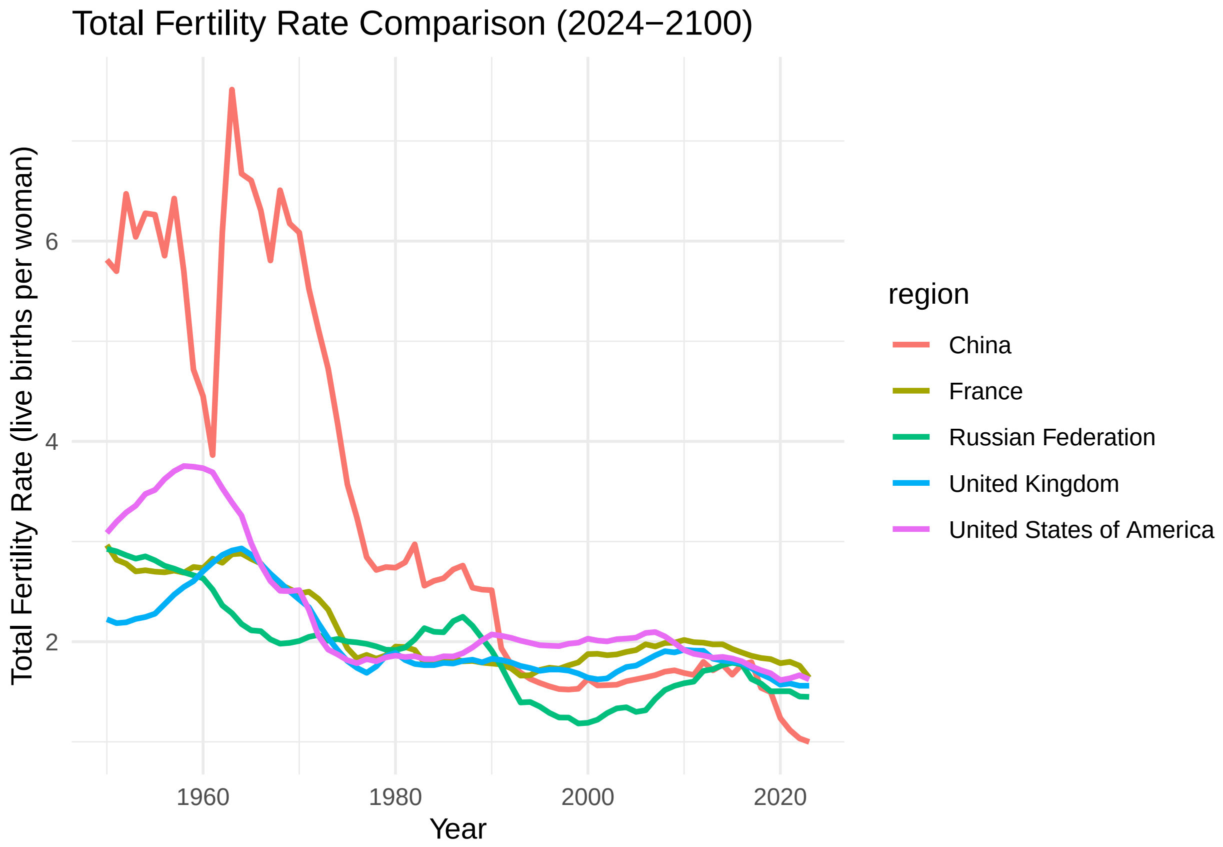
Explanation 6 This bar chart shows the population densities, or population per square km, of the five global super powers being studied. It was interesting to see the United Kingdom dominating in this category especially with China having a much higher overall population than the other four countries. This graph highlights the urbanized population of the United Kingdom and limited land area that comes with it, compared to overall population. Meanwhile we can see Russia with a very low population density, highlighting their extremely large landmass.

```
ggplot(global_superpowers_data, aes(x = year, y = total_deaths, fill = region)) +
  geom_area(alpha = 0.7) +
  labs(
    title = "Total Deaths Over Time for Global Superpowers",
    x = "Year",
    y = "Total Deaths",
    fill = "Country"
  ) +
  theme_minimal()
```



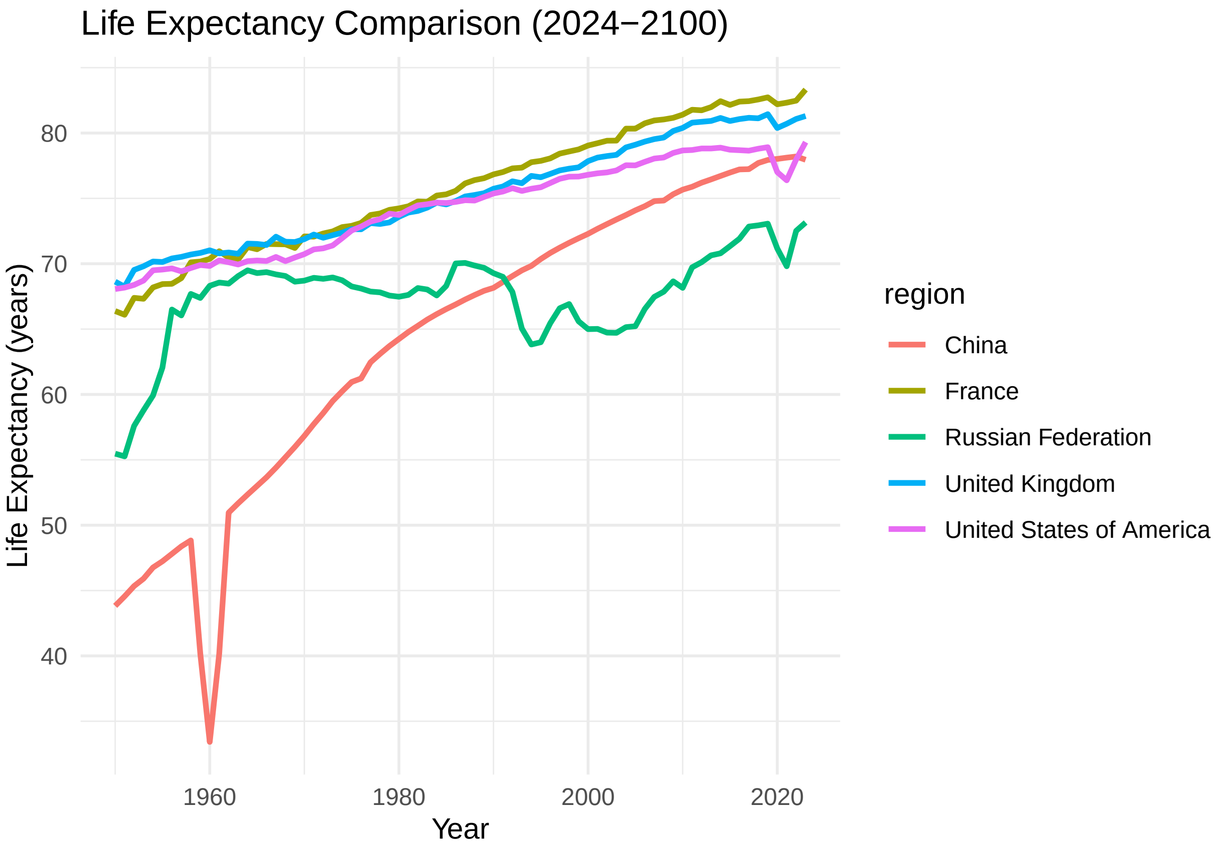
Explanation 7 This area chart shows the distribution of total deaths between the five countries. With China having a much higher population and population density than than the other four countries given in other graphs, it makes sense for China to have a much higher proportion of deaths than the other two, as illustrated in the pie chart above.

```
ggplot(global_superpowers_data, aes(x = year, y = total_fertility_rate, color = region)) +
geom_line(size = 1) +
labs(
title = "Total Fertility Rate Comparison (2024-2100)",
x = "Year",
y = "Total Fertility Rate (live births per woman)") + theme_minimal()
```



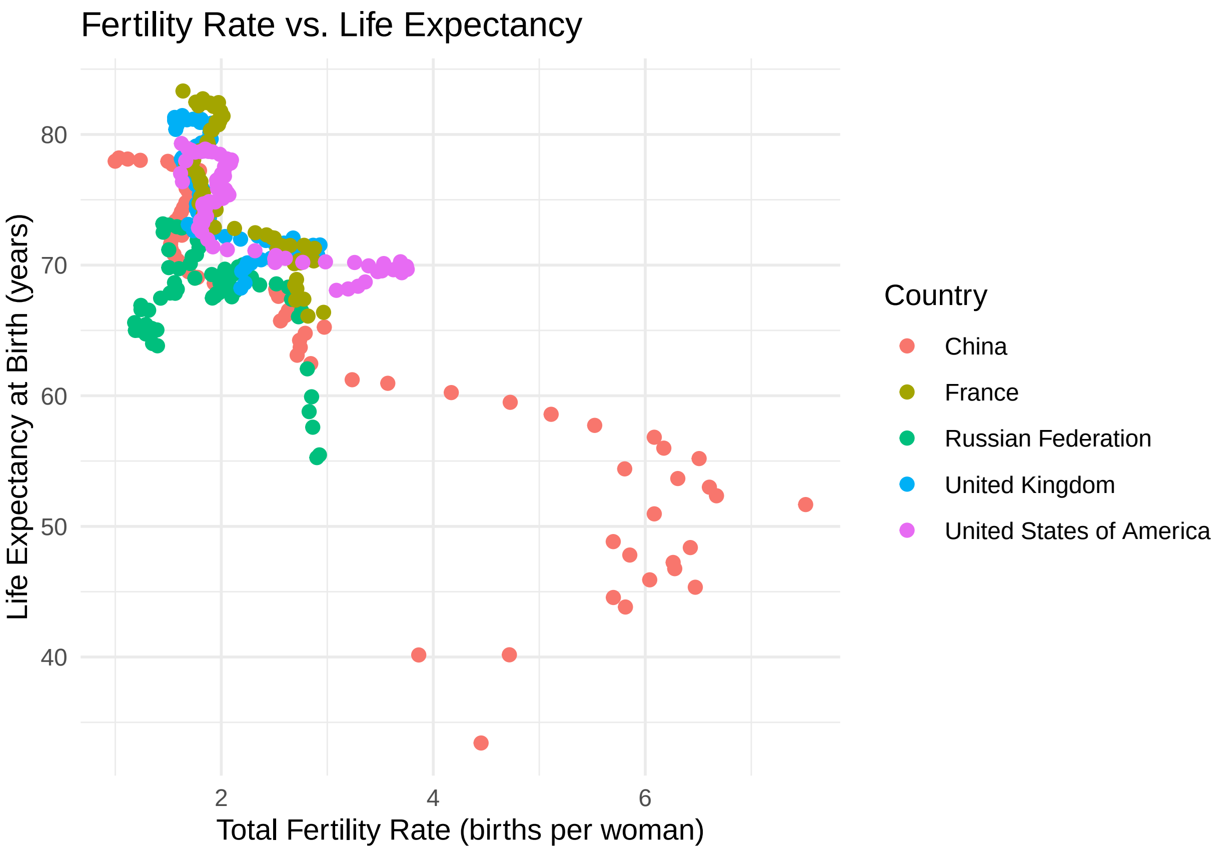
Explanation 8 This line chart examines the fertility rate of the five countries over time, in births per women. All five countries show an overall decreasing trend as time goes on, relating to our difference in fertility map, which showed a negative number for all five countries as well. We continue to see a sharp decline and largest decrease in China's fertility rate which could very well be in response too cultural shifts and its long-standing population control policies. China takes a massive dip around the 1960s, when the Mao's great famine took place, lowering the birth rate.

```
ggplot(global_superpowers_data, aes(x = year, y = life_exp_birth_both, color = region)) +
geom_line(size = 1) +
labs(
title = "Life Expectancy Comparison (2024-2100)",
x = "Year",
y = "Life Expectancy (years)") +theme_minimal()
```



Explanation 9 This graph shows us the average life expectancy between the three countries. Once again, China shows an interesting slope here with a drastic negative decrease in 1960, caused by the famine known as Mao's Great famine, lowering life expectancy greatly. yet overall, China clearly shows the steepest increasing slope, all three countries show an overall increase in life expectancy over time. This makes sense as the world in general has made advancements in living conditions, nutrition, medicine and more. United States, United Kingdom, France and China show somewhat similar growth while Russia shows slower. These disparities in growth can be due to factors such as healthcare systems and limitations over time, which would lead to less life expectancy growth.

```
ggplot(global_superpowers_data, aes(x = total_fertility_rate, y = life_exp_birth_both, color = region))
  geom_point(size = 2) +
  labs(
    title = "Fertility Rate vs. Life Expectancy",
    x = "Total Fertility Rate (births per woman)",
    y = "Life Expectancy at Birth (years)",
    color = "Country"
  ) +
  theme_minimal()
```

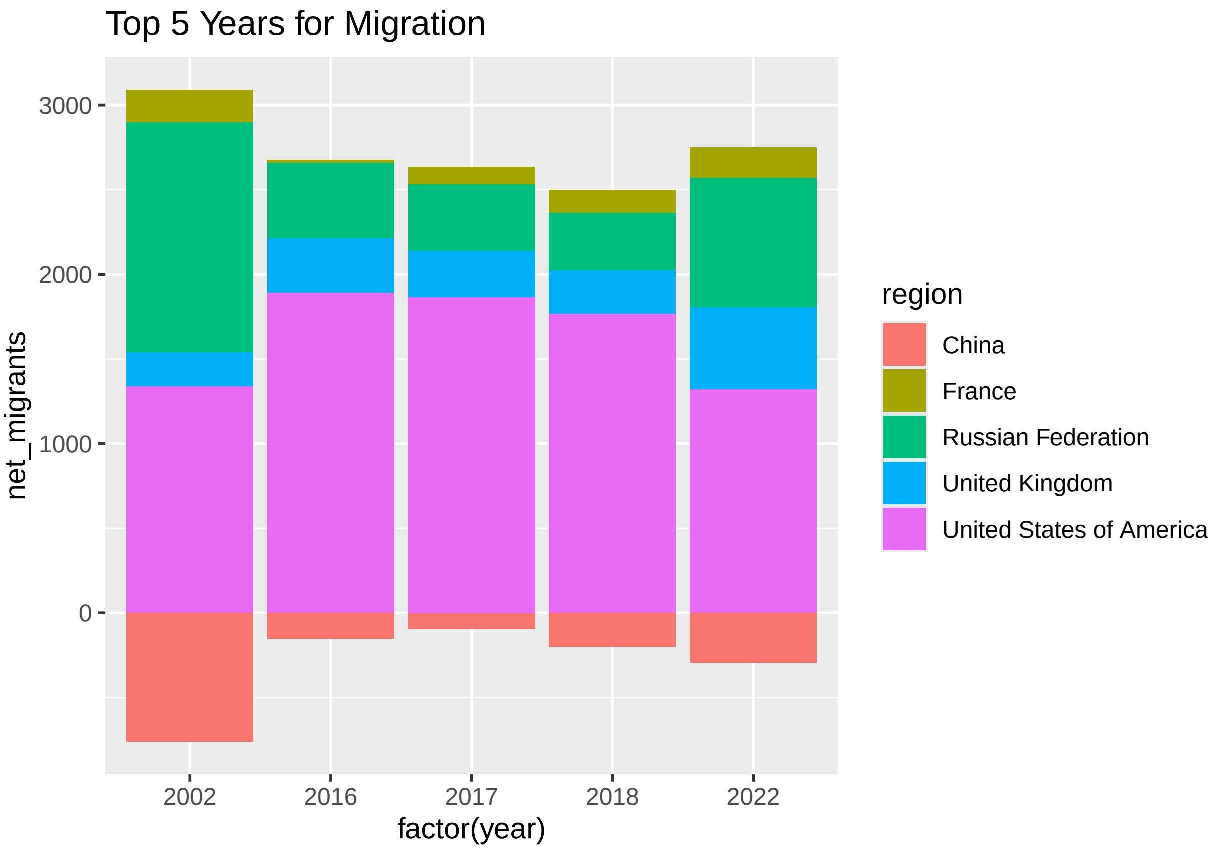


Explanation 10 This plot illustrates the relationship between total fertility rate and life expectancy for the global superpower countries. While these plots do not show the strongest and most clear correlation, all five countries show an inverse relationship in general. In other words, lower fertility rates tend to correlate to greater life expectancy while higher fertility rates tend to lead towards lower life expectancies.

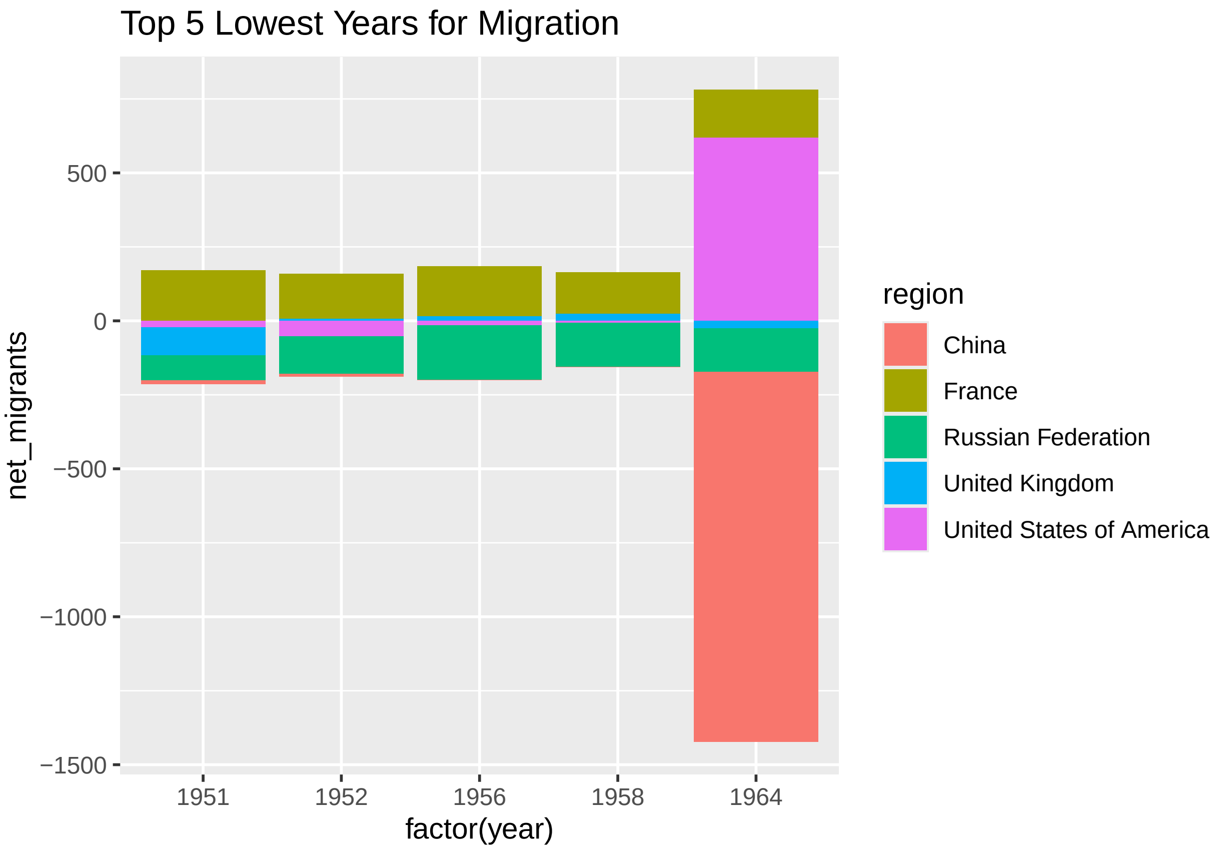
```
migrant_data <- estimates_df |>
  filter(region %in% countries) |>
  select(region, year, net_migrants) |>
  group_by(year) |>
  reframe(total = sum(net_migrants), across())
```

```
## Warning: There was 1 warning in `reframe()`.
## i In argument: `across()`.
## Caused by warning:
## ! Using `across()` without supplying `.cols` was deprecated in dplyr 1.1.0.
## i Please supply `.cols` instead.
```

```
migrant_data |> arrange(desc(total)) |>
  ungroup() |>
  slice_head(n = 25) |>
  ggplot(aes(x = factor(year), y = net_migrants, fill = region)) +
  geom_col() +
  labs(title = "Top 5 Years for Migration")
```

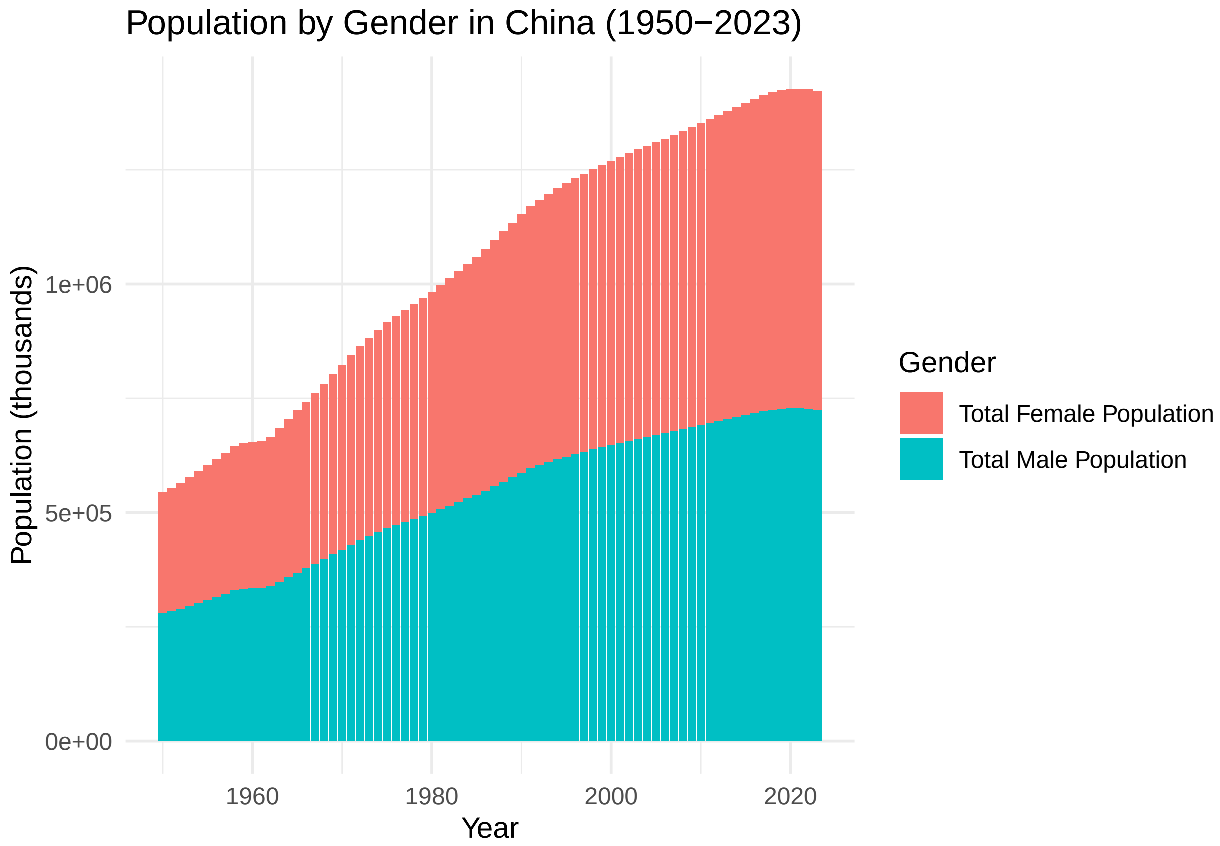


```
migrant_data |> arrange(total) |>
  ungroup() |>
  slice_head(n = 25) |>
  ggplot(aes(x = factor(year), y = net_migrants, fill = region)) +
  geom_col() +
  labs(title = "Top 5 Lowest Years for Migration")
```



Explanation 11 When comparing the top 5 years of migration vs the the lowest 5 years of migration, we can see that the top 5 years are primarily within the 21st century, while the top lowest years are in the 50's and 60's. A likely explanation for the lack of migration during that time is the Cold War: With global superpowers at odds, this likely decreasing the desire to migrate away from / into these countries. Migration is high during the 2000's likely because of globalization. As economies became more intertwined, this likely lead to workers and families moving for job and other economical opportunities.

```
age_data <- global_superpowers_data %>%
  filter(region == "China") %>%
  select(year, male_pop_july, female_pop_july) %>%
  pivot_longer(
    cols = c(male_pop_july, female_pop_july),
    names_to = "Gender",
    values_to = "Population"
  ) %>%
  mutate(Gender = recode(Gender,
                         male_pop_july = "Total Male Population",
                         female_pop_july = "Total Female Population"))

ggplot(age_data, aes(x = year, y = Population, fill = Gender)) +
  geom_bar(stat = "identity", position = "stack") +
  labs(
    title = "Population by Gender in China (1950-2023)",
    x = "Year",
    y = "Population (thousands)",
    fill = "Gender"
  ) +
  theme_minimal()
```



Explanation 12 This plot depicts the gender distribution in China from 1950 to 2023. Although the proportions may not be perfectly precise, the data clearly show that the female population has gradually caught up with the male population over time. Specifically, the total female population has moved closer to a 50-50 balance compared to the male population. This trend indicates a more balanced gender distribution in China's population as the years progress.

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.

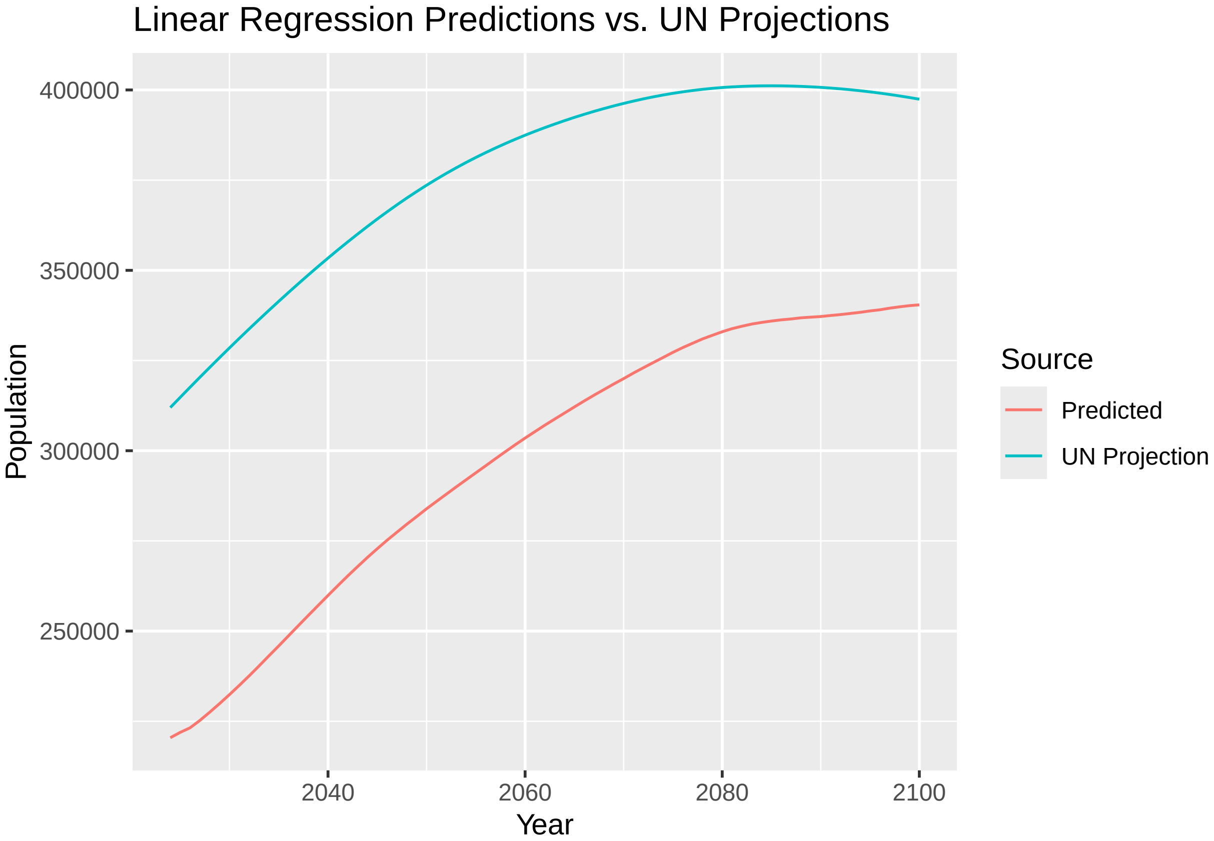
The UN's World Population Prospects provides a range of projected scenarios of population change. These rely on different assumptions in fertility, mortality and/or migration patterns to explore different demographic futures. Check this link for more info: https://population.un.org/wpp/DefinitionOfProjectionScenarios

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.

```
medium_scenario_proj <- estimates_df |> select(year, total_pop_july, births, total_deaths, pop_change)

proj_linear_model <- lm(total_pop_july ~ births + total_deaths + pop_change, data = medium_scenario_pro

new_prediction <- predict(proj_linear_model, newdata = medium_variant_df)

compare_model <- medium_variant_df |>
```

```
  select(year, total_pop_july) |>
  mutate(predicted_population = new_prediction) |> na.omit()

compare_model_graph <- compare_model |> group_by(year) |> summarise(avg_UN_pop = mean(total_pop_july),

ggplot(compare_model_graph, aes(x = year, y = total_pop_july)) +
  geom_line(aes(y = avg_model_pop, color = "Predicted")) +
  geom_line(aes(y = avg_UN_pop, color = "UN Projection")) +
  labs(
    title = "Linear Regression Predictions vs. UN Projections",
    x = "Year",
    y = "Population",
    color = "Source"
  )
```



Explanation Our models predictions are significantly lower than that of the UN projection. While we maintain a very similar trend as the UN projection, illustrating that there is some accuracy and similarity between the two models, our prediction is much lower. This may suggest that the UN model is factoring in some other variables that are making the population predictions higher on average. **6. Requirement-5 (1 pt)**

Conclusion

Some of the trends we discovered in our analysis included seeing a steady decline in fertility rate for almost all countries included in the analysis. While this may be due to circumstances like geo-political conflict, we also suspect that this phenomenon could be due to cultural factors, such as modern feminism encouraging women globally to prioritize work over home-making. We also notice that migration to the UK, the Russian Federation, and the United States dropped significantly in 2016 as compared to previous years, when migration was steadily increasing. This may be the result of the pandemic, which led to countries implementing stricter immigration policies as a precaution. Though life expectancy for all countries in the analysis dropped at the time of the pandemic, it seems to have recovered at about the same rate for all countries. It was also very interesting to see a reoccurring drastic change in China's graphs around the 1960 time period. For instance population and fertility rate were taking a huge dip while death made a huge spike. These changes were likely due to the Great Chinese Famine, or Mao's Great Famine, that occurred as a result of the country's attempt to rapidly industrialize and modernize agriculture.

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