APEC 8221 - Assignment 1: Tidyverse Fundamentals

[Yunshu Yang]

2025-09-15

GitHub Repository URL: [Insert your repository URL here]

1 Setup

```
# Load required packages
library(tidyverse)
library(gapminder)

# Check data structure
glimpse(gapminder)
```

```
Rows: 1,704
Columns: 6
$ country   <fct> "Afghanistan", "Afghanistan", "Afghanistan", "Afghanistan", ~
$ continent <fct> Asia, Asia, Asia, Asia, Asia, Asia, Asia, Asia, Asia, Asia, ~
$ year      <int> 1952, 1957, 1962, 1967, 1972, 1977, 1982, 1987, 1992, 1997, ~
$ lifeExp   <dbl> 28.801, 30.332, 31.997, 34.020, 36.088, 38.438, 39.854, 40.8~
$ pop       <int> 8425333, 9240934, 10267083, 11537966, 13079460, 14880372, 12~
$ gdpPercap <dbl> 779.4453, 820.8530, 853.1007, 836.1971, 739.9811, 786.1134, ~
```

2 Data Analysis & Visualization Practice

2.1 Top Performers Identification

Find the top 4.3 countries in 2.each continent for 3.highest life expectancy and highest GDP per capita in 1.2007.

2.1.1 Highest Life Expectancy by Continent (2007)

```
# Your code here - find top 3 countries by life expectancy in each continent
# Remember to use group_by(), arrange(), and slice_head()
gapminder %>%
  filter(year == 2007) %>%              # Step 1: keep 2007 data only
  group_by(continent) %>%               # Step 2: group data by continent
  arrange(desc(lifeExp)) %>%            # Step 3: sort by life expectancy (high to low)
  slice_head(n = 3) %>%                 # Step 4: select the top 3 per group
  select(continent, country, lifeExp)   # Step 5: keep only needed columns
```

```
# A tibble: 14 x 3
# Groups:   continent [5]
   continent country          lifeExp
   <fct>     <fct>              <dbl>
 1 Africa    Reunion             76.4
 2 Africa    Libya               74.0
 3 Africa    Tunisia             73.9
 4 Americas  Canada              80.7
 5 Americas  Costa Rica          78.8
 6 Americas  Puerto Rico         78.7
 7 Asia      Japan               82.6
 8 Asia      Hong Kong, China    82.2
 9 Asia      Israel              80.7
10 Europe    Iceland             81.8
11 Europe    Switzerland         81.7
12 Europe    Spain               80.9
13 Oceania   Australia           81.2
14 Oceania   New Zealand         80.2
```

2.1.2 Highest GDP per Capita by Continent (2007)

```
# Your code here - find top 3 countries by GDP per capita in each continent
gapminder %>%
  filter(year == 2007) %>%                # Step 1: keep 2007 data
  group_by(continent) %>%                 # Step 2: group by continent
  arrange(desc(gdpPercap)) %>%            # Step 3: sort by GDP per capita (high to low)
  slice_head(n = 3) %>%                   # Step 4: keep top 3 per group
  select(continent, country, gdpPercap)   # Step 5: display relevant columns
```

```
# A tibble: 14 x 3
# Groups:   continent [5]
   continent country           gdpPercap
   <fct>     <fct>                 <dbl>
 1 Africa    Gabon                13206.
 2 Africa    Botswana             12570.
 3 Africa    Equatorial Guinea    12154.
 4 Americas  United States        42952.
 5 Americas  Canada               36319.
 6 Americas  Puerto Rico          19329.
 7 Asia      Kuwait               47307.
 8 Asia      Singapore            47143.
 9 Asia      Hong Kong, China     39725.
10 Europe    Norway               49357.
11 Europe    Ireland              40676.
12 Europe    Switzerland          37506.
13 Oceania   Australia            34435.
14 Oceania   New Zealand          25185.
```

Brief analysis: *[Write 2-3 sentences about what patterns you notice in the top performers]*: In 2007, countries with the highest life expectancy were mostly high-income economies such as Japan, Iceland, and Australia. Interestingly, the countries with the highest GDP per capita (e.g., Norway, Kuwait, and Singapore) were not always the ones with the highest life expectancy. This suggests that while wealth is strongly associated with better health outcomes, other social, healthcare, and policy factors also play important roles in determining life expectancy.

2.2 Development Progress Visualization

Create a scatter plot comparing each country's 1952 vs. 2007 life expectancy.

```
# Step 1: Create separate data frames for 1952 and 2007
  # Your code here - filter for 1952 and select relevant columns
lifeExp_1952 <- gapminder %>%
  filter(year == 1952) %>%
  select(country, lifeExp_1952 = lifeExp)

  # Your code here - filter for 2007 and select relevant columns (include population!)
lifeExp_2007 <- gapminder %>%
  filter(year == 2007) %>%
  select(country, continent, pop, lifeExp_2007 = lifeExp) #continents are the same as 1952 a
```

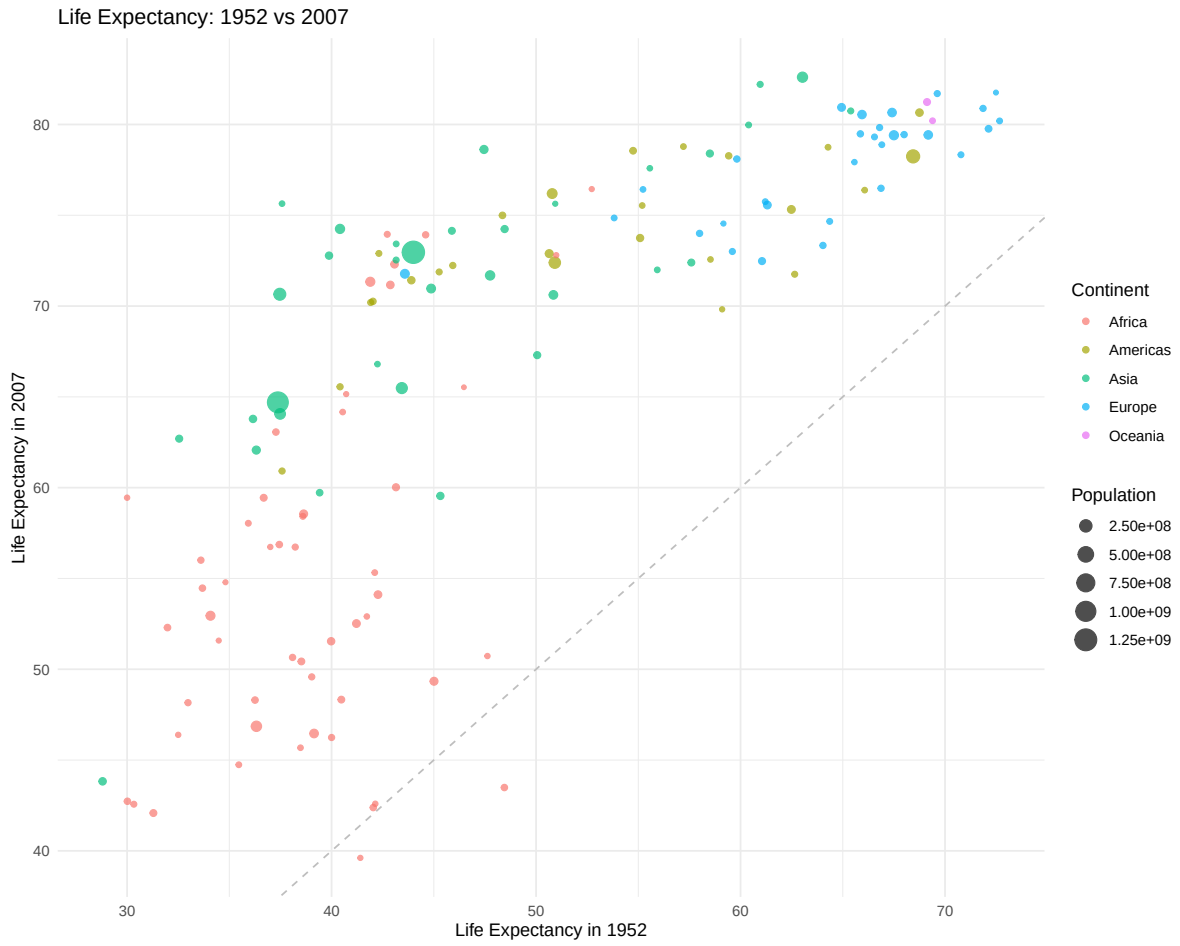
```r
# Step 2: Combine using provided left_join code
lifeExp_data <- left_join(lifeExp_1952, lifeExp_2007, by = "country")

# Step 3: Handle missing data
lifeExp_data <- lifeExp_data %>%
  # Your code here - remove countries with missing data using filter()
 filter(!is.na(lifeExp_1952), !is.na(lifeExp_2007))

#check
head(lifeExp_data)
```

```
# A tibble: 6 x 5
  country     lifeExp_1952 continent      pop lifeExp_2007
  <fct>              <dbl> <fct>        <int>        <dbl>
1 Afghanistan         28.8 Asia      31889923         43.8
2 Albania             55.2 Europe     3600523         76.4
3 Algeria             43.1 Africa    33333216         72.3
4 Angola              30.0 Africa    12420476         42.7
5 Argentina           62.5 Americas  40301927         75.3
6 Australia           69.1 Oceania   20434176         81.2
```

```r
# Create the scatter plot
ggplot(lifeExp_data, aes(x = lifeExp_1952, y = lifeExp_2007)) +
  # Your code here - add the diagonal reference line using geom_abline()
  # Add points with color by continent and size by population
  # Add professional labels and theme
geom_abline(slope = 1, intercept = 0, linetype = "dashed", color = "gray") +
  geom_point(aes(color = continent, size = pop), alpha = 0.7) +
  labs(
    title = "Life Expectancy: 1952 vs 2007",
    x = "Life Expectancy in 1952",
    y = "Life Expectancy in 2007",
    color = "Continent",
    size = "Population"
  ) +
  theme_minimal()
```



Analysis: *[Write 2-3 sentences explaining what the plot shows about global health progress]* : The scatter plot shows that nearly all countries lie above the diagonal line, meaning life expectancy in 2007 was consistently higher than in 1952. The size of the points highlights how improvements in large-population countries such as China and India had a major influence on global health progress. At the same time, differences remain across regions: while Europe and Oceania reached the highest life expectancies, many African countries still lagged behind, showing persistent global health disparities.

2.3 Development Progress Classification

Using the `lifeExp_data` from Task 2.1.2, classify countries by improvement level.

```
# Add classification to your existing lifeExp_data
lifeExp_data <- lifeExp_data %>%
  mutate(
```

```
    improvement = lifeExp_2007 - lifeExp_1952,   # years gained
    improvement_level = case_when(
      improvement >= 30 ~ "Major Improvement",
      improvement >= 15 ~ "Moderate Improvement",
      improvement >= 0  ~ "Minor Improvement",
      improvement < 0   ~ "Decline"
    )
  )

# Create summary table
  # Your code here - group by continent and improvement_level, count countries
improvement_summary <- lifeExp_data %>%
  group_by(continent, improvement_level) %>%
  summarise(n_countries = n(), .groups = "drop")


print(improvement_summary)
```

```
# A tibble: 13 x 3
   continent improvement_level    n_countries
   <fct>     <chr>                      <int>
 1 Africa    Decline                        2
 2 Africa    Major Improvement              1
 3 Africa    Minor Improvement             23
 4 Africa    Moderate Improvement          26
 5 Americas  Major Improvement              1
 6 Americas  Minor Improvement              8
 7 Americas  Moderate Improvement          16
 8 Asia      Major Improvement              7
 9 Asia      Minor Improvement              2
10 Asia      Moderate Improvement          24
11 Europe    Minor Improvement             23
12 Europe    Moderate Improvement           7
13 Oceania   Minor Improvement              2
```

2.3.1 Analysis Questions

1. Which continent has the most countries with "Major Improvement"?

```r
# Your code here - analyze major improvement by continent
# Count countries with Major Improvement by continent
major_improvement <- lifeExp_data %>%
  filter(improvement_level == "Major Improvement") %>%
  group_by(continent) %>%
  summarise(n_countries = n(), .groups = "drop") %>%
  arrange(desc(n_countries))

print(major_improvement)
```

```
# A tibble: 3 x 2
  continent n_countries
  <fct>           <int>
1 Asia                7
2 Africa              1
3 Americas            1
```

[Your answer here]: Asia has the most countries with "Major Improvement," with 7 countries showing a gain of 30 or more years in life expectancy between 1952 and 2007. In contrast, only one country in Africa and one in the Americas fall into this category, highlighting Asia's substantial health progress during this period.

2. Are there any countries that experienced decline? Which ones and why might this have happened?

```r
# Your code here - find and examine countries with decline
# Find countries with decline
decline_countries <- lifeExp_data %>%
  filter(improvement_level == "Decline") %>%
  select(country, continent, lifeExp_1952, lifeExp_2007, improvement)

print(decline_countries)
```

```
# A tibble: 2 x 5
  country   continent lifeExp_1952 lifeExp_2007 improvement
  <fct>     <fct>            <dbl>        <dbl>       <dbl>
1 Swaziland Africa            41.4         39.6       -1.79
2 Zimbabwe  Africa            48.5         43.5       -4.96
```

[Your answer here]: Yes, two countries experienced a decline in life expectancy between 1952 and 2007: Swaziland (Eswatini) and Zimbabwe, both located in Africa. These declines are

largely attributed to the HIV/AIDS epidemic, which severely reduced life expectancy in southern Africa during the 1990s and 2000s. In addition, Zimbabwe also faced political and economic crises that weakened its healthcare system and contributed to worsening health outcomes.

3. What does this tell us about global health convergence?

[Your answer here - 2-3 sentences about convergence patterns]: The results suggest partial convergence in global health: while nearly all countries improved their life expectancy between 1952 and 2007, the magnitude of improvement varied widely across regions. Many Asian countries achieved major gains, while some African countries experienced stagnation or even decline, showing that global health progress is uneven and convergence is incomplete.

2.4 Economic Development Trajectories

Create a line plot showing life expectancy changes over time for selected countries.

2.4.1 Step 1: Guided Discovery

```
# Find interesting countries for your plot
  # Your code here - calculate change and identify top/bottom performers
life_exp_changes <- lifeExp_data %>%
  mutate(change = lifeExp_2007 - lifeExp_1952) %>%
  arrange(desc(change))

# Print top gainers and any decliners
top_gainers <- life_exp_changes %>%
  slice_max(change, n = 5) %>%
  select(country, continent, lifeExp_1952, lifeExp_2007, change)
print(top_gainers)
```

```
# A tibble: 5 x 5
  country      continent lifeExp_1952 lifeExp_2007 change
  <fct>        <fct>            <dbl>        <dbl>  <dbl>
1 Oman         Asia              37.6         75.6   38.1
2 Vietnam      Asia              40.4         74.2   33.8
3 Indonesia    Asia              37.5         70.6   33.2
4 Saudi Arabia Asia              39.9         72.8   32.9
5 Libya        Africa            42.7         74.0   31.2
```

```r
decliners <- life_exp_changes %>%
  filter(change < 0) %>%
  select(country, continent, lifeExp_1952, lifeExp_2007, change)
print(decliners)
```

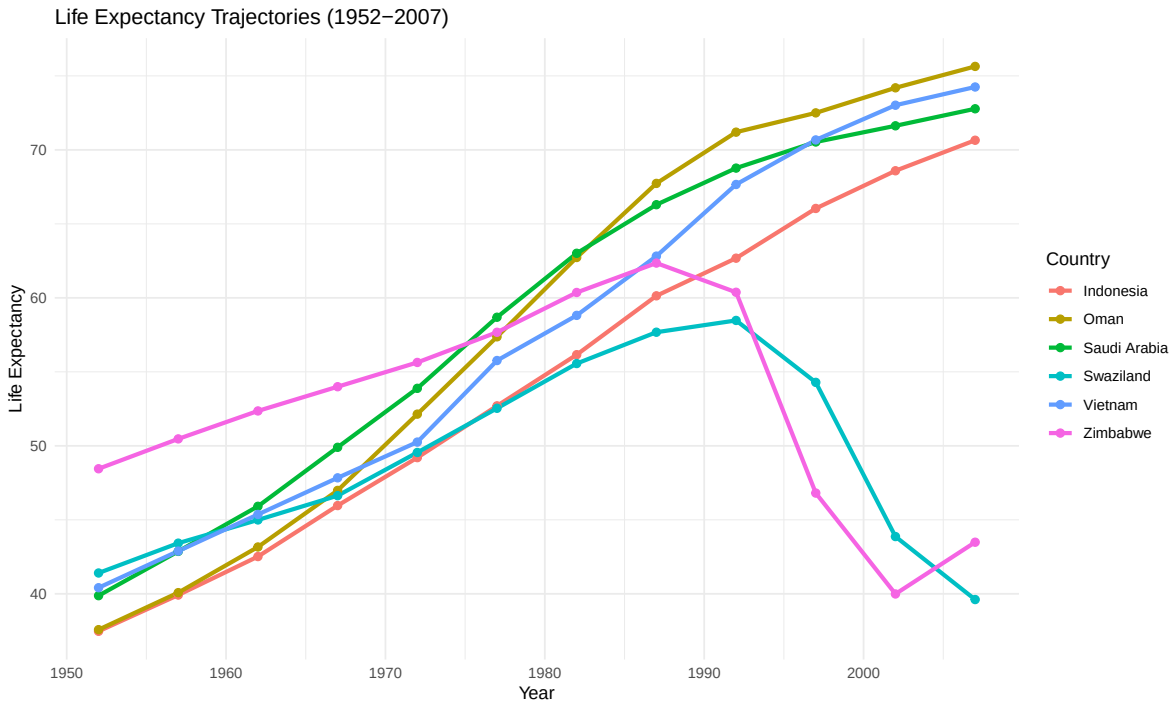
```
# A tibble: 2 x 5
  country   continent lifeExp_1952 lifeExp_2007 change
  <fct>     <fct>            <dbl>        <dbl>  <dbl>
1 Swaziland Africa            41.4         39.6  -1.79
2 Zimbabwe  Africa            48.5         43.5  -4.96
```

2.4.2 Step 2: Create Trajectory Visualization

```r
# Select 5-6 interesting countries based on your discovery
selected_countries <- c("Oman", "Vietnam", "Indonesia",
                        "Saudi Arabia", "Swaziland", "Zimbabwe")
  # Your countries here - include top gainer, any decliner, and 3-4 others of interest


  # Your code here - filter for selected countries
trajectory_data <- gapminder %>%
  filter(country %in% selected_countries)


# Create line plot
ggplot(trajectory_data, aes(x = year, y = lifeExp, color = country)) +
  geom_line(linewidth = 1.2) +
  geom_point(size = 2) +
  labs(
    title = "Life Expectancy Trajectories (1952-2007)",
    x = "Year",
    y = "Life Expectancy",
    color = "Country"
  ) +
  theme_minimal()
```




```
# Your code here - add lines, points, and professional labeling
```

Caption: *[Write 2-3 sentences explaining why you selected these countries and what pattern stands out]*: These trajectories highlight striking contrasts in global health development. Oman, Vietnam, Saudi Arabia, and Indonesia show dramatic improvements in life expectancy, reflecting rapid social and economic progress. In contrast, Swaziland and Zimbabwe experienced declines after the 1990s due to the HIV/AIDS epidemic, illustrating how health crises can reverse decades of progress.

3 Research Discovery & Critical Analysis

3.1 Choose Your Research Question

Selected Research Question: *[A, B, C, D, or E]* **Which countries achieved double improvement (GDP and life expectancy both increased) between 1952 and 2007, and which ones did not? What patterns can we observe across these groups?**

3.2 Research Analysis

3.2.1 Data Analysis

```
# Your code here - use appropriate dplyr operations to investigate your question
# This should include filtering, grouping, summarizing, and/or calculating as needed

# Step 1: Extract 1952 and 2007 data
gap_1952 <- gapminder %>%
  filter(year == 1952) %>%
  select(country, gdpPercap_1952 = gdpPercap, lifeExp_1952 = lifeExp)

gap_2007 <- gapminder %>%
  filter(year == 2007) %>%
  select(country, continent, gdpPercap_2007 = gdpPercap, lifeExp_2007 = lifeExp)

# Step 2: Join and calculate changes
gap_change <- left_join(gap_1952, gap_2007, by = "country") %>%
  select(country, continent, gdpPercap_1952, gdpPercap_2007,
         lifeExp_1952, lifeExp_2007) %>%
  mutate(
    gdp_change = gdpPercap_2007 - gdpPercap_1952,
    lifeExp_change = lifeExp_2007 - lifeExp_1952,
    double_improve = ifelse(gdp_change > 0 & lifeExp_change > 0, "Yes", "No")
  )

#list double improvement
double_yes <- gap_change %>%
  filter(double_improve == "Yes") %>%
  select(country, continent, gdpPercap_1952, gdpPercap_2007, gdp_change,
         lifeExp_1952, lifeExp_2007, lifeExp_change)

print(double_yes)
```

```
# A tibble: 128 x 8
   country     continent gdpPercap_1952 gdpPercap_2007 gdp_change lifeExp_1952
   <fct>       <fct>              <dbl>          <dbl>      <dbl>        <dbl>
 1 Afghanistan Asia                779.           975.       195.         28.8
 2 Albania     Europe             1601.          5937.      4336.         55.2
 3 Algeria     Africa             2449.          6223.      3774.         43.1
 4 Angola      Africa             3521.          4797.      1277.         30.0
```

```
 5 Argentina   Americas          5911.         12779.      6868.         62.5
 6 Australia   Oceania          10040.         34435.     24396.         69.1
 7 Austria     Europe            6137.         36126.     29989.         66.8
 8 Bahrain     Asia              9867.         29796.     19929.         50.9
 9 Bangladesh  Asia               684.          1391.       707.         37.5
10 Belgium     Europe            8343.         33693.     25349.         68
# i 118 more rows
# i 2 more variables: lifeExp_2007 <dbl>, lifeExp_change <dbl>
```

```r
#list double non-improvement
double_no <- gap_change %>%
  filter(double_improve == "No") %>%
  select(country, continent, gdpPercap_1952, gdpPercap_2007, gdp_change,
         lifeExp_1952, lifeExp_2007, lifeExp_change)

print(double_no)
```

```
# A tibble: 14 x 8
   country       continent gdpPercap_1952 gdpPercap_2007 gdp_change lifeExp_1952
   <fct>         <fct>              <dbl>          <dbl>      <dbl>        <dbl>
 1 Central Afri~ Africa             1071.          706.     -365.          35.5
 2 Comoros       Africa             1103.          986.     -117.          40.7
 3 Congo, Dem. ~ Africa              781.          278.     -503.          39.1
 4 Djibouti      Africa             2670.         2082.     -587.          34.8
 5 Haiti         Americas           1840.         1202.     -639.          37.6
 6 Kuwait        Asia             108382.        47307.   -61075.          55.6
 7 Liberia       Africa              576.          415.     -161.          38.5
 8 Madagascar    Africa             1443.         1045.     -398.          36.7
 9 Nicaragua     Americas           3112.         2749.     -363.          42.3
10 Niger         Africa              762.          620.     -142.          37.4
11 Sierra Leone  Africa              880.          863.      -17.2         30.3
12 Somalia       Africa             1136.          926.     -210.          33.0
13 Swaziland     Africa             1148.         4513.     3365.          41.4
14 Zimbabwe      Africa              407.          470.       62.8         48.5
# i 2 more variables: lifeExp_2007 <dbl>, lifeExp_change <dbl>
```

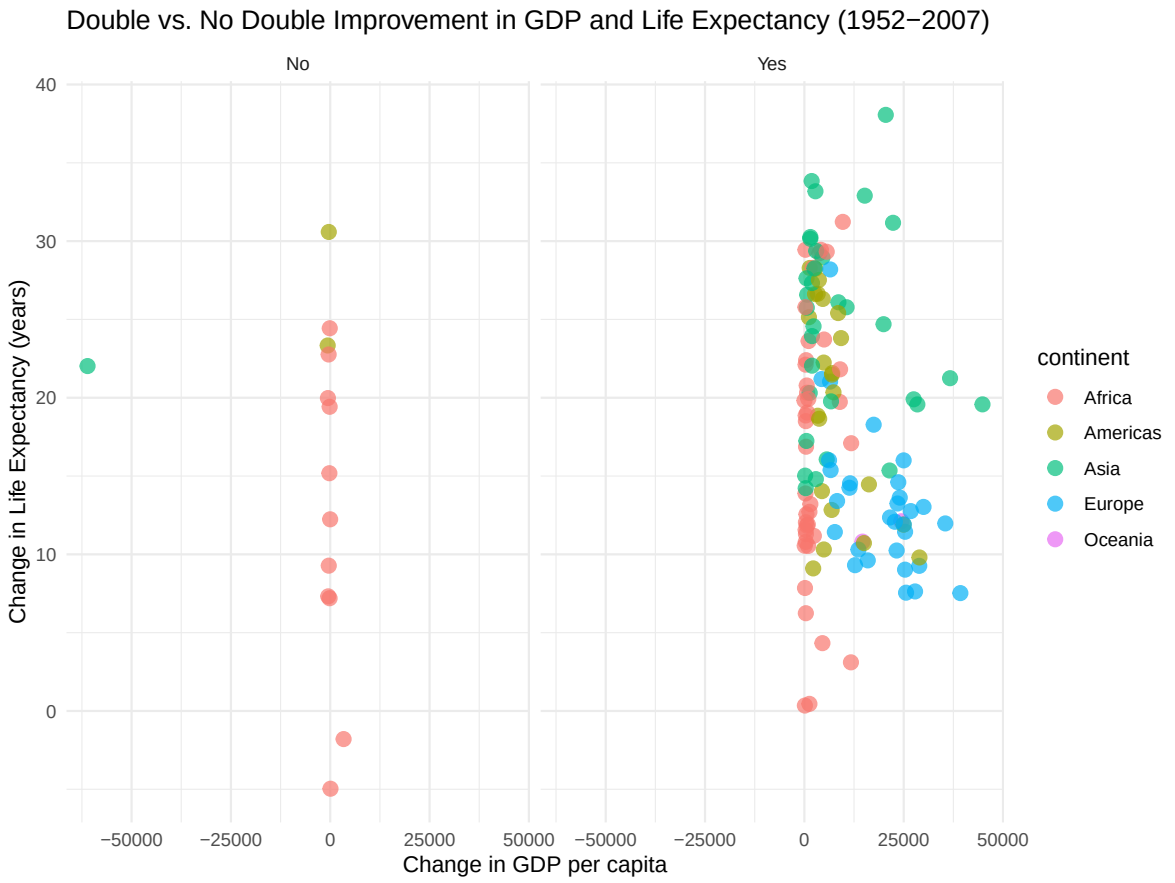
```r
#summary
gap_change %>%
  group_by(double_improve, continent) %>%
  summarise(n_countries = n(),
            avg_gdp_change = mean(gdp_change, na.rm = TRUE),
```

```
            avg_lifeExp_change = mean(lifeExp_change, na.rm = TRUE),
            .groups = "drop")
```

```
# A tibble: 8 x 5
  double_improve continent n_countries avg_gdp_change avg_lifeExp_change
  <chr>          <fct>           <int>          <dbl>               <dbl>
1 No             Africa             11           84.3                11.9
2 No             Americas            2         -501.                 27.0
3 No             Asia                1       -61075.                 22.0
4 Yes            Africa             41         2307.                 16.7
5 Yes            Americas           23         7570.                 19.8
6 Yes            Asia               32         9414.                 24.5
7 Yes            Europe             30        19393.                 13.2
8 Yes            Oceania             2        19512.                 11.5
```

3.2.2 Visualization

```
# Your code here - create one clear plot that supports your findings
ggplot(gap_change, aes(x = gdp_change, y = lifeExp_change, color = continent)) +
  geom_point(size = 3, alpha = 0.7) +
  facet_wrap(~ double_improve) +   #seperate
 labs(
  title = "Double vs. No Double Improvement in GDP and Life Expectancy (1952-2007)",
  x = "Change in GDP per capita",
  y = "Change in Life Expectancy (years)"
) +
  theme_minimal()
```



3.2.3 Interpretation (150-200 words)

[Your interpretation here - explain what you found, why it's interesting, potential explanations for the pattern, and any limitations of your analysis]: This analysis explored whether countries experienced "double improvement," defined as simultaneous gains in both GDP per capita and life expectancy between 1952 and 2007. The majority of countries fell into the "double improvement" category, highlighting a broad pattern of economic growth accompanied by health progress during this period. However, a small subset of countries did not show this dual progress. For example, several African countries saw improvements in life expectancy but little or negative GDP growth, while a few Asian and African countries showed economic stagnation or decline.

These findings are interesting because they reveal that economic development and population health often move together, but not universally. Contextual factors such as political stability, governance, public health interventions, and the HIV/AIDS epidemic in sub-Saharan Africa may help explain why some countries failed to experience improvements in both dimensions.

One limitation of this analysis is that it relies only on two time points (1952 and 2007) and cannot capture short-term fluctuations or regional crises. Additionally, GDP per capita and life expectancy, while widely used, are imperfect measures of overall well-being and may obscure within-country inequality.