APEC 8221 - Assignment 1: Tidyverse Fundamentals

Due Saturday, September 20, 2025 at 8:00 AM

Table of contents

1 Assignment Overview

1.1 Dataset

We'll continue using the `gapminder` dataset for this assignment, which contains temporal (every 5 years) information on life expectancy, population and GDP per capita for 142 countries worldwide:

- `country`: Country name
- `continent`: Africa, Americas, Asia, Europe, Oceania

- `year`: 1952, 1957, 1962, 1967, 1972, 1977, 1982, 1987, 1992, 1997, 2002, 2007
- `lifeExp`: Life expectancy at birth (years)
- `pop`: Total population
- `gdpPercap`: GDP per capita (US$, inflation-adjusted)

1.2 Technical Requirements

- Use the provided `assignment1-template.qmd` Quarto template
- Save your file as `assignment1-[your-x500-id].qmd`
- All file paths must be relative (R Project structure)
- Code must be well-commented with clear section headers
- Quarto document must knit to PDF without errors
- Include your GitHub repository URL in your submission so I can review your Quarto document
- Git history should show logical progression with meaningful commit messages

1.3 Grading Rubric

Criteria	Needs Improvement	Satisfactory	Excellent	Points
Correctness & Accuracy	Code fails to run or produces incorrect results	Code runs with largely correct results, minor errors possible	Code is correct, efficient, and performs all tasks accurately	20 pts
Code Clarity & Style	Poor formatting, lacks comments, unclear logic	Readable code with adequate comments	Clean, well-commented code following style guidelines	8 pts
Reproducibility & Quarto	Document won't knit or uses absolute paths	Document knits successfully with relative paths	Professional document with clear methodology	6 pts
Version Control	Single commit or no meaningful history	Several commits with generic messages	Clear development story with descriptive commits	6 pts

2 Tasks

2.1 Data Analysis & Visualization Practice

2.1.1 Top Performers Identification

Find the **top 3 countries** in each continent for:

1. Highest life expectancy (2007)
2. Highest GDP per capita (2007)

Write 2-3 sentences about what patterns you notice in the top performers.

Requirements:

- Use `group_by()`, `arrange()`, and `slice_head()`
- Present results in two separate, well-formatted tables
- Include continent, country, and the relevant metric

2.1.2 Development Progress Visualization

Create a **professional scatter plot** comparing each country's **1952 vs. 2007 life expectancy**, with:

- Points colored by continent
- Point size proportional to 2007 population
- A diagonal reference line (y = x) showing "no change" [Hint: `?geom_abline`]
- Countries above the line improved; below the line declined
- Professional labels, legend, and theme

Requirements:

- Create separate data frames for 1952 (`lifeExp_1952`) and 2007 (`lifeExp_2007`). Be sure to include population in your 2007 data frame.
- Since we haven't covered joins yet, you can use the provided code below to combine your two data frames:

```
# Combine the datasets (we'll learn joins in Week 4)
lifeExp_data <- left_join(lifeExp_1952, lifeExp_2007, by = "country")
```

- Filter the data to only countries that have data for both 1952 and 2007.

Removing missing observations

```
filter(!is.na(variable1) & !is.na(variable2))
```

- Write 2-3 sentences explaining what the plot shows about global health progress.

2.1.3 Development Progress Classification

Using the `lifeExp_data` you created in Task 2.1.2, classify countries by their life expectancy improvement and analyze regional patterns.

Requirements:

1. Create improvement categories using the life expectancy data you already prepared:
    1. **Major Improvement**: 30+ years gained
    2. **Moderate Improvement**: 15-29 years gained
    3. **Minor Improvement**: 0-14 years gained
    4. **Decline**: Lost life expectancy

Creating classifications

```
# For simple two-category classification:
mutate(category = if_else(condition, "Category A", "Category B"))

# For multiple categories, use case_when():
mutate(
  category = case_when(
    condition1 ~ "Category A",
    condition2 ~ "Category B",
    condition3 ~ "Category C",
    TRUE ~ "Category D"  # catch-all for remaining cases
  )
)
```

2. Create a summary table showing how many countries in each continent fall into each improvement category.

Note

Do not worry about creating a formal "table"—printing your data frame is sufficient. We'll discuss publication-ready tables later in the class

3. Analysis questions:

    1. Which continent has the most countries with "Major Improvement"?
    2. Are there any countries that experienced decline? Which ones and why might this have happened?
    3. What does this tell us about global health convergence?

2.1.4 Economic Development Trajectories

Create a line plot showing how life expectancy has changed over time for selected countries.

Step 1: Guided Discovery First, explore the data to identify:

- Which country had the highest life expectancy gain between 1952-2007?
- Which country had the lowest life expectancy gain (or biggest decline)?
- Which country's trajectory surprises you most when you look at the data?

Step 2: Create Your Visualization Create a line plot using your discoveries from Step 1 plus 2-3 additional countries of your choice (total of 5-6 countries):

- Year on x-axis, life expectancy on y-axis
- One line per country, colored by country name for clear distinction
- Add points at each data point with `geom_point()`
- Clear legend showing country names Informative title that highlights your key finding
- Professional labels, title, and theme
- Include a brief caption (2-3 sentences) explaining why you selected these countries

2.2 Research Discovery & Critical Analysis

2.2.1 Choose Your Research Question

Select **one** of the following research questions to investigate. Each uses only the skills we've learned so far:

A. **The Middle-Income Trap** Identify countries that achieved middle-income status (GDP per capita $3,000-$12,000) by 1987 but failed to reach high-income status by 2007. What patterns do you observe?

B. **Initial Wealth and Development Paradox** First, explore the 1952 GDP per capita distribution to identify reasonable high/low thresholds (or use top quartile vs. bottom quartile). [Hint: `?quantile()`] Then compare how these different initial wealth groups performed in terms of life expectancy improvements 1952-2007. Do countries that started wealthy show different development trajectories?

C. **Demographic Dividend Analysis** Identify countries that experienced rapid population growth (>100% increase 1952-2007). Did this help or hurt their economic development per capita?

D. **Continental Development Patterns** Compare how different continents have progressed over time. Which continent shows the most improvement in life expectancy? Which shows the most economic growth? Are these the same?

E. **Your Own Question** Design your own research question using only `filter()`, `select()`, `mutate()`, `summarise()`, `group_by()`, `arrange()`, and basic `ggplot2` functions on the `gapminder` data.

2.2.2 Research Requirements

For your chosen question:

1. **Conduct Analysis:** Use appropriate `dplyr` operations to investigate your question
2. **Create Visualization:** Design one clear plot that supports your findings
3. **Write Interpretation:** Provide 150-200 words explaining:
    - What you found and why it's interesting
    - Potential explanations for the pattern
    - Any limitations of your analysis