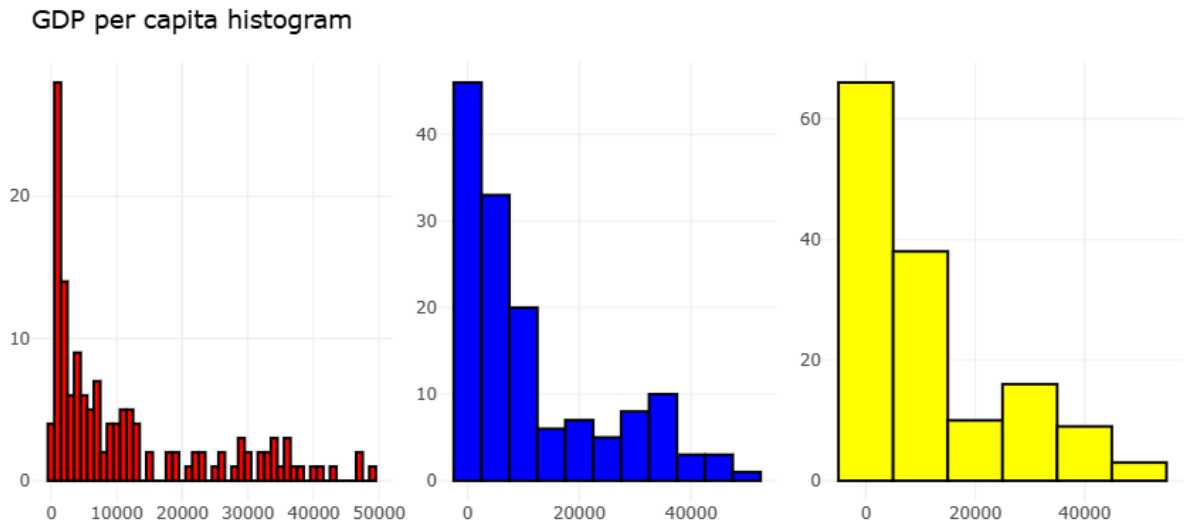
Rishita Agarwal - 220150016

DA332
Lab Assignment 5

Link to the notebook - https://www.kaggle.com/code/rishita26/da-332-lab-5

Q1)

a)



1. The choice of bin width significantly affects how we interpret the GDP per capita distribution:

Bin width = 1000: This narrow bin width creates many bins, revealing fine details in the distribution. It clearly shows that most countries have GDP per capita below 10,000, with a long right tail. You can see small peaks and valleys that might indicate clusters of countries at particular income levels.

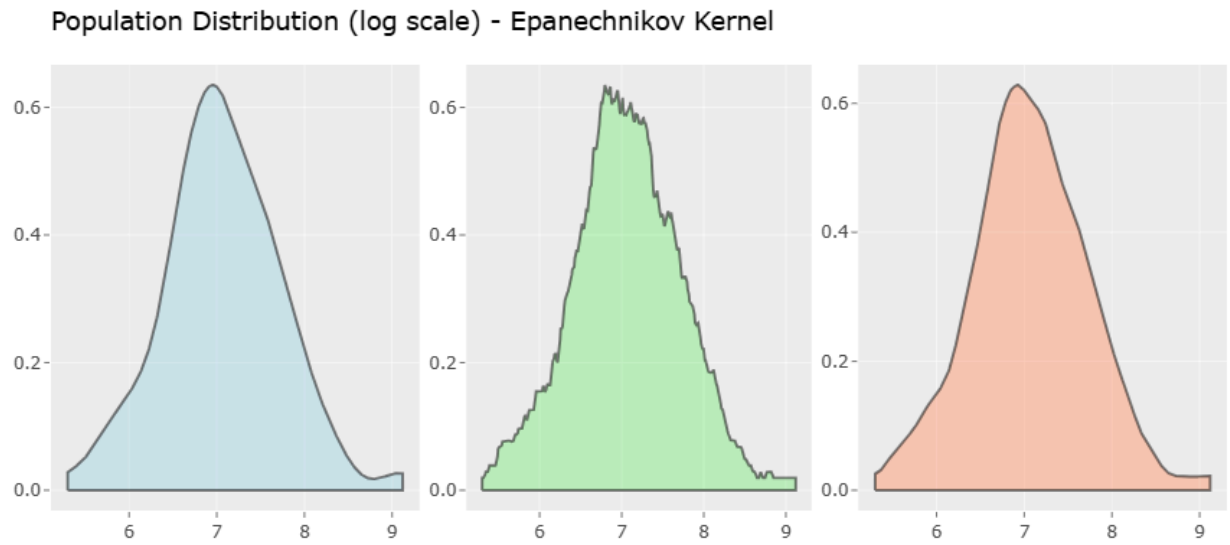
Bin width = 5000: This medium bin width provides a balance between detail and overall shape. The distribution still shows the skewness toward lower values, but some of the fine details are smoothed out. The overall pattern becomes more apparent.

Bin width = 10000: This wide bin width greatly simplifies the distribution, showing only the broadest features. The extreme right skew is evident, but specific clusters within the data are no longer visible as they get combined into fewer, wider bins.

2. The bin width of **5000** provides the most meaningful insight for this dataset because:

- It balances detail and clarity - showing the overall shape without being too noisy
- It clearly demonstrates the right-skewed nature of global GDP distribution
- It still reveals important clusters (like the concentration of countries with GDP per capita below 10,000)
- It's not too granular (like 1000) where random variation might be mistaken for patterns
- It's not too broad (like 10000) where important features might be lost

b)



The three kernel functions affect the population density estimation differently:

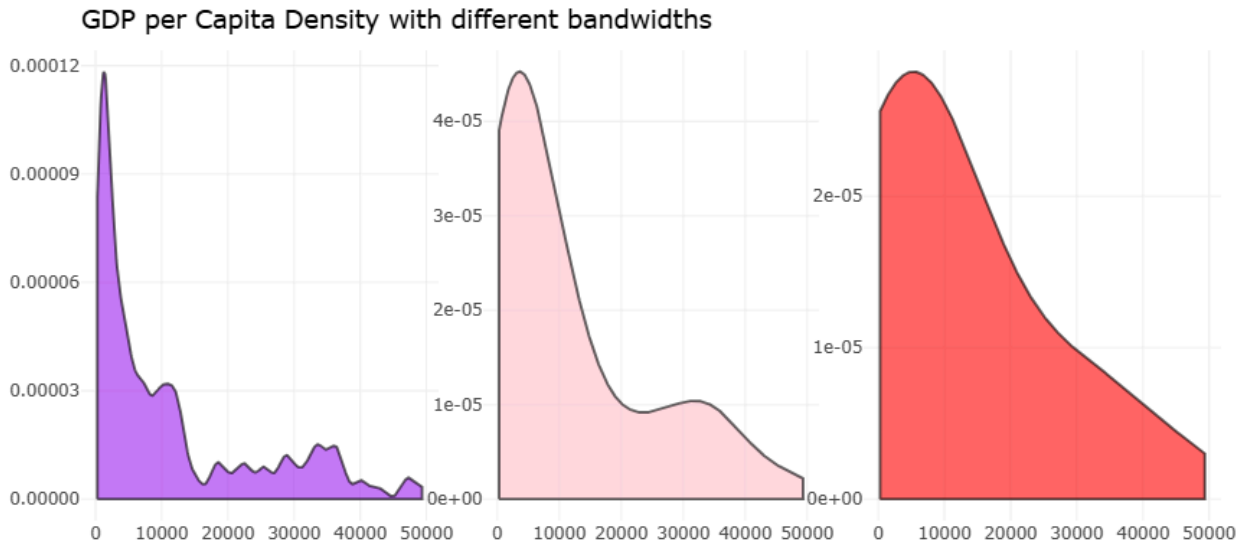
Gaussian kernel: Produces the smoothest curve with gradually tapering tails. It shows a relatively smooth bimodal distribution in the population data, suggesting two clusters of countries by population size. The transitions between different parts of the distribution are gradual.

Rectangular kernel: Creates a more jagged, step-like appearance with abrupt changes. The bimodal structure is still visible, but the transitions between density regions are more sudden. This kernel preserves more localized features but can look artificial at times.

Epanechnikov kernel: Offers a middle ground between the other two kernels. It maintains a relatively smooth appearance with less extreme tails than the Gaussian. The bimodal structure is clearly visible, with a balance of smoothness and preservation of data features.

Each kernel makes different assumptions about how to weigh nearby observations, affecting how the density is estimated around each data point. The Epanechnikov kernel is often considered optimal in statistical theory as it minimizes asymptotic mean integrated squared error.

c)



The bandwidth parameter acts as a smoothing control for density plots:

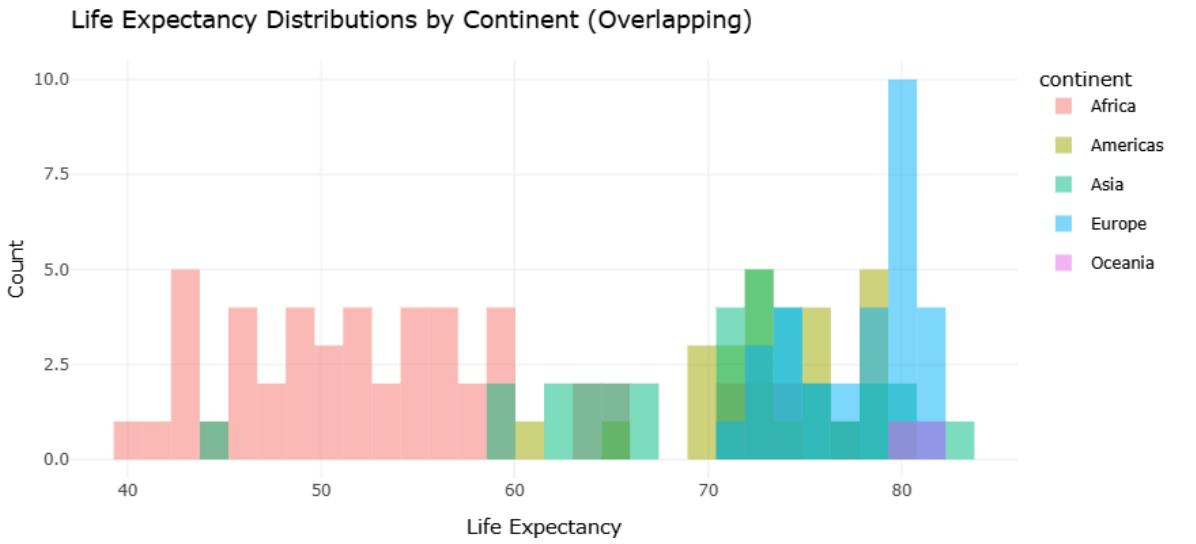
Bandwidth = 1000: This narrow bandwidth creates a density plot with high detail, showing multiple peaks and local variations. It reveals fine structures in the data but may overfit to random noise, potentially showing patterns that aren't meaningful.

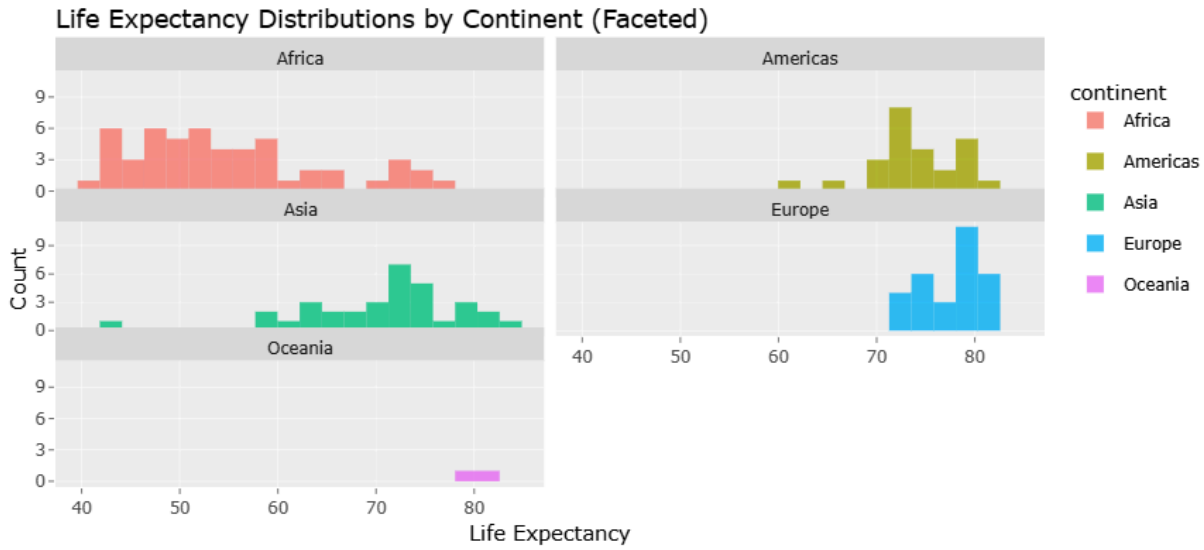
Bandwidth = 5000: This medium bandwidth strikes a balance between detail and smoothness. The major features of the distribution remain visible (like the concentration of lower GDP values), but minor fluctuations are smoothed out. This often provides a more realistic representation of the underlying distribution.

Bandwidth = 10000: This wide bandwidth creates a very smooth curve that highlights only the broadest features of the distribution. It clearly shows the right-skewed nature of global GDP distribution but masks potential multimodality or clusters that might exist in the data.

Bandwidth selection is crucial because too small a bandwidth can result in a "spiky" density that overfits to the sample, while too large a bandwidth can over smooth and hide important features of the data.

d)



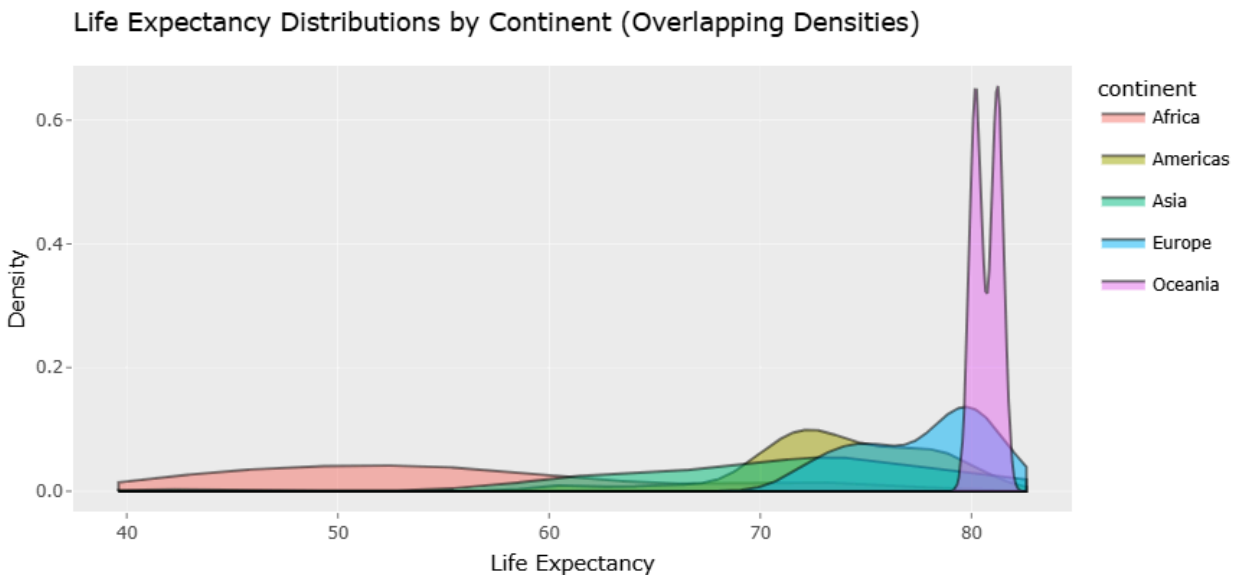


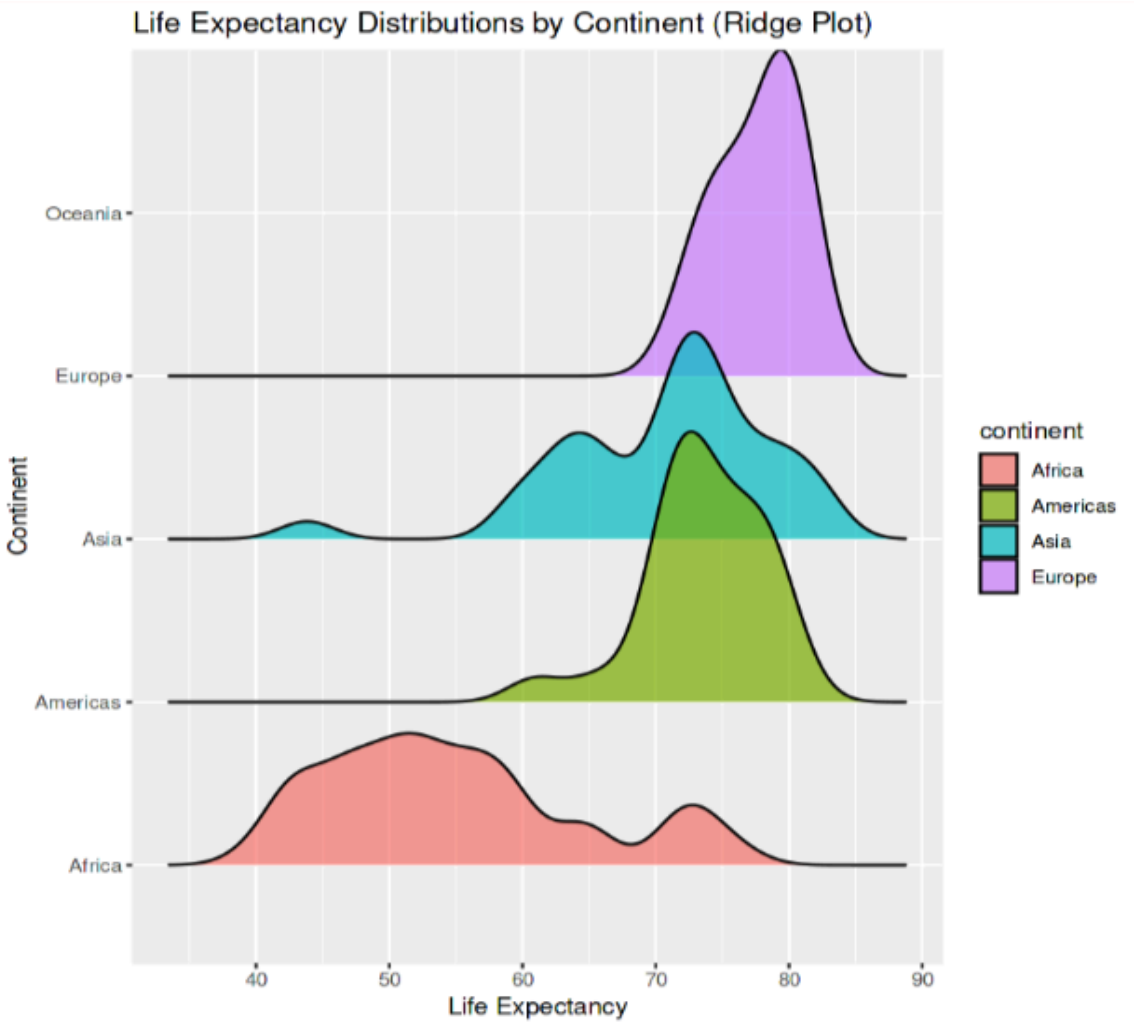
The faceted approach is better for comparing life expectancy distributions across continents for several reasons:

- Clarity: Each continent gets its own separate panel, eliminating the visual confusion caused by overlapping distributions.
- Equal visual weight: In the overlapping histogram, continents with fewer countries (like Oceania) are barely visible, while in the faceted approach, each continent gets equal visual space.
- Accurate comparisons: The faceted approach allows viewers to accurately compare the shapes of distributions without the distortion caused by overlapping bars.
- Scale adaptability: Each facet can have its own y-axis scale, allowing better visualization of the distribution shape for continents with fewer countries.
- Pattern recognition: The distinct separation makes it easier to identify unique patterns for each continent (e.g., Africa's bimodal distribution is much clearer in the faceted view).

While the overlapping histogram does provide a direct comparison of the ranges, the visual clutter significantly hampers interpretation of the actual distributions.

e)



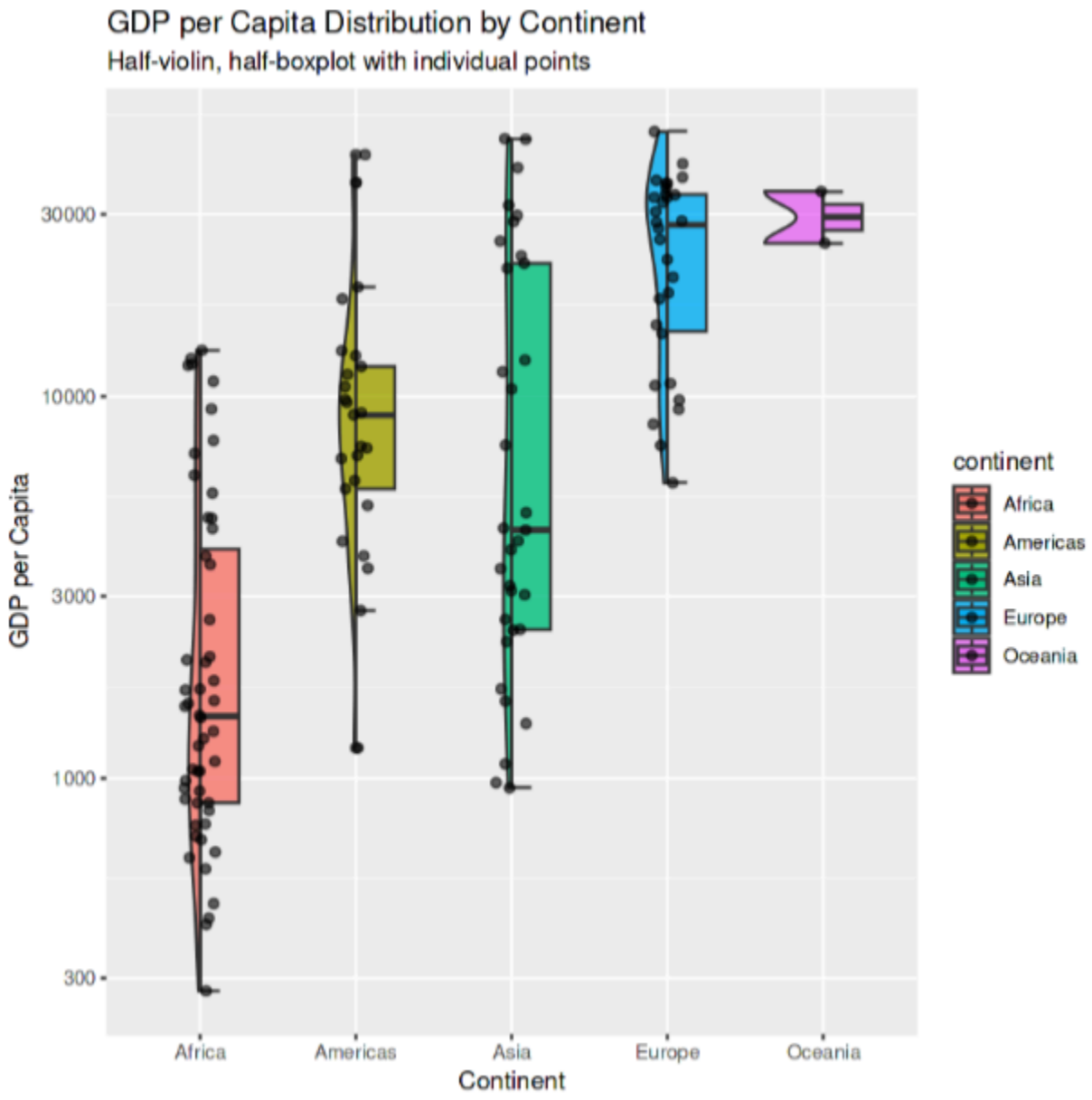


Ridge plots are superior to overlapping density plots for several reasons:

- Reduced occlusion: Ridge plots eliminate the problem of overlapping distributions hiding key features. In the overlapping densities plot, it's difficult to see the full shape of distributions that share the same x-range.
- Ordered comparison: Ridge plots arrange continents in a meaningful vertical order, making it easier to compare medians and ranges across continents.
- Shape preservation: Each distribution's complete shape is visible, allowing better visualization of important features like bimodality (particularly evident in Africa's distribution).
- Visual hierarchy: Ridge plots create a natural visual hierarchy that draws attention to the differences between continents rather than differences within continents.
- Aesthetic appeal: Ridge plots are easier to read and visually more appealing, making the insights more accessible to viewers.

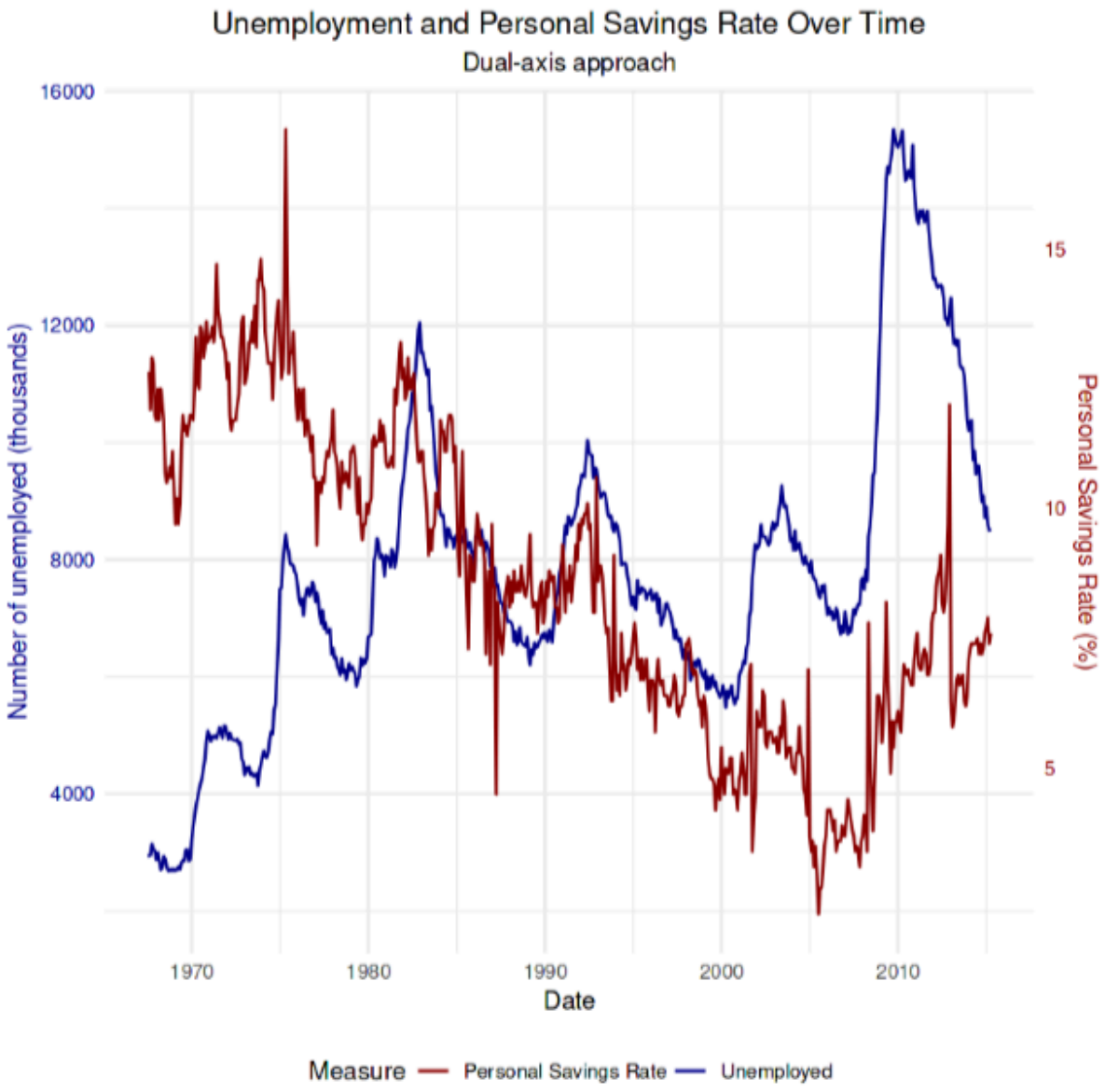
The ridge plot clearly shows, for example, that Africa has a bimodal distribution of life expectancy (suggesting two distinct groups of countries), while Europe has a tight, unimodal distribution centered at higher life expectancies.

f)

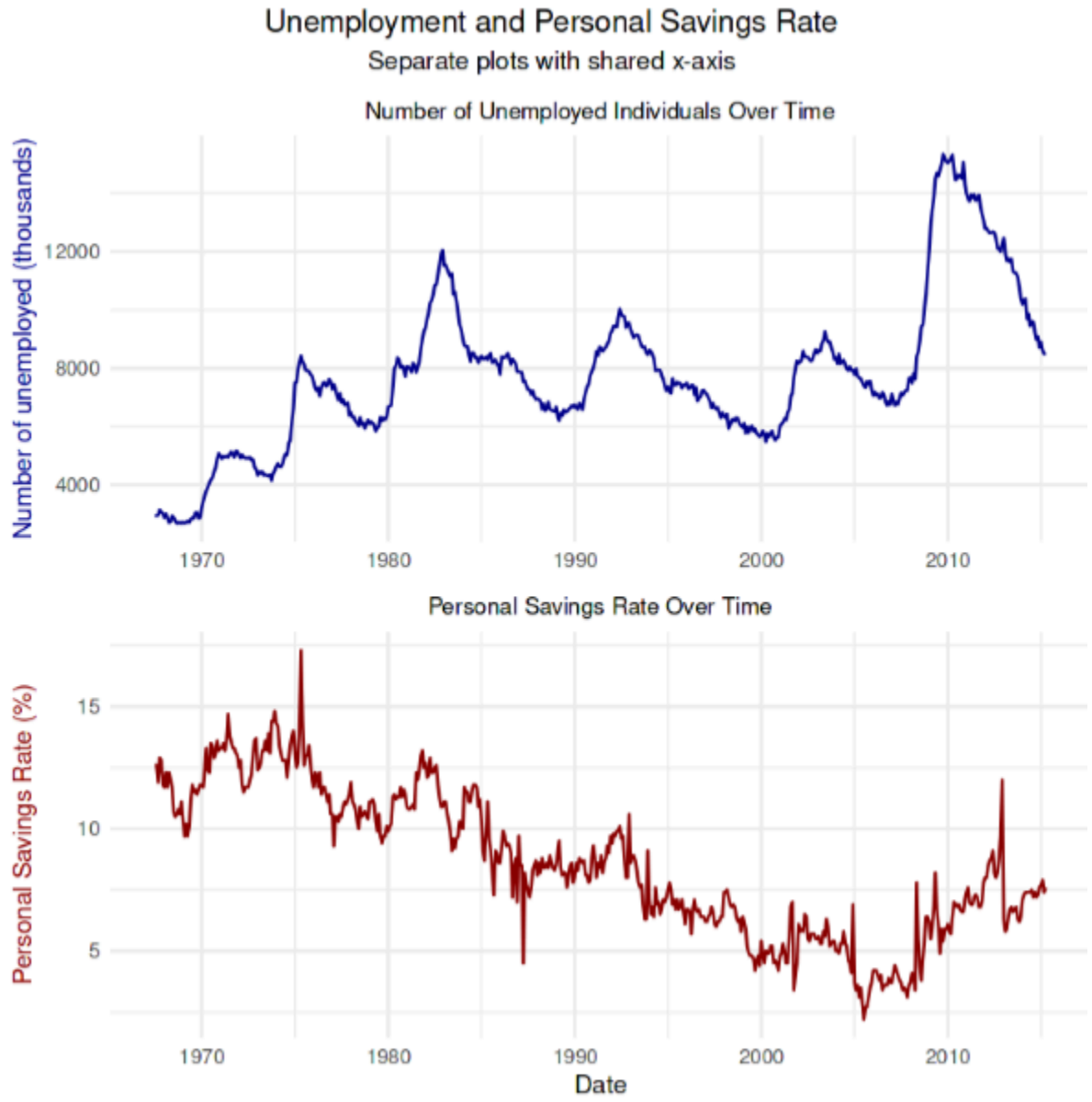

GDP per Capita Distribution by Continent
Half-violin, half-boxplot with individual points
GDP per Capita
30000
10000
3000
1000
300
Africa
Americas
Asia
Europe
Oceania
Continent
continent
Africa
Americas
Asia
Europe
Oceania

Q2)

a)


Unemployment and Personal Savings Rate Over Time
Dual-axis approach
16000
12000
8000
4000
Number of unemployed (thousands)
15
10
5
Personal Savings Rate (%)
1970
1980
1990
2000
2010
Date
Measure
Personal Savings Rate
Unemployed

b)



c)

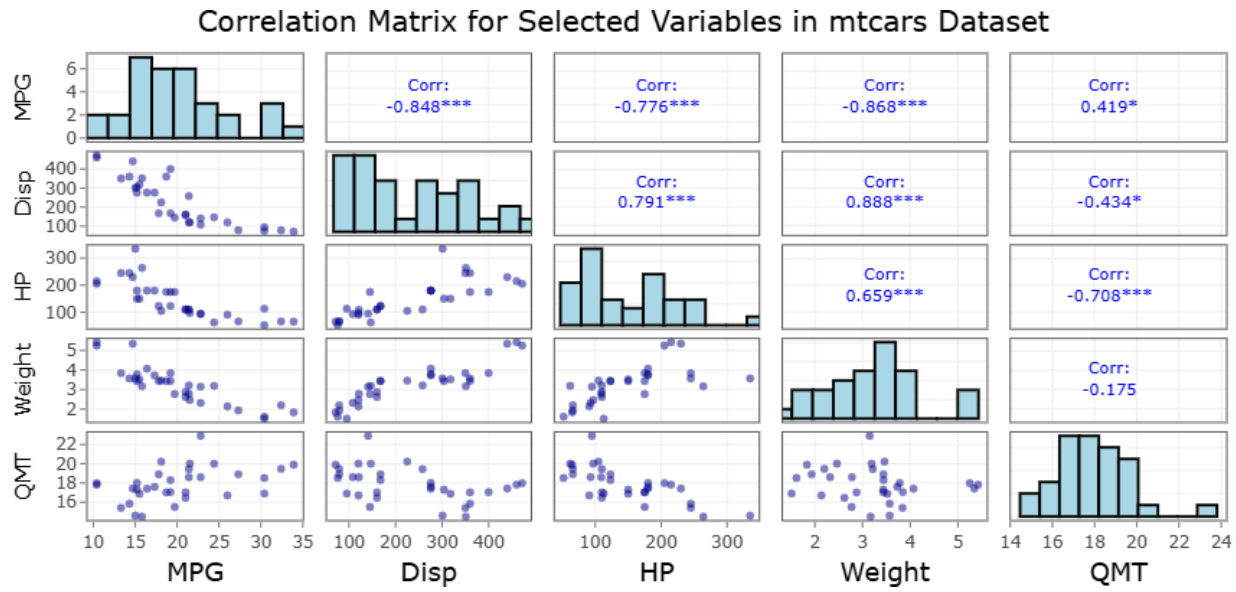
Using patchwork is better than the first approach because :

- Avoids Misleading Visual Correlations: Dual-axis plots can create the illusion of correlation between variables when none exists. The scaling process required to fit both variables on one plot can distort the viewer's perception of the relationship between them.
- Preserves Data Integrity: Each variable is plotted on its own scale without manipulation, maintaining the integrity of the original data. In the dual-axis plot, one variable had to be artificially rescaled.
- Improves Readability: Having separate plots eliminates confusion about which line corresponds to which axis. Even with color coding, dual-axis plots require more cognitive effort to interpret correctly.

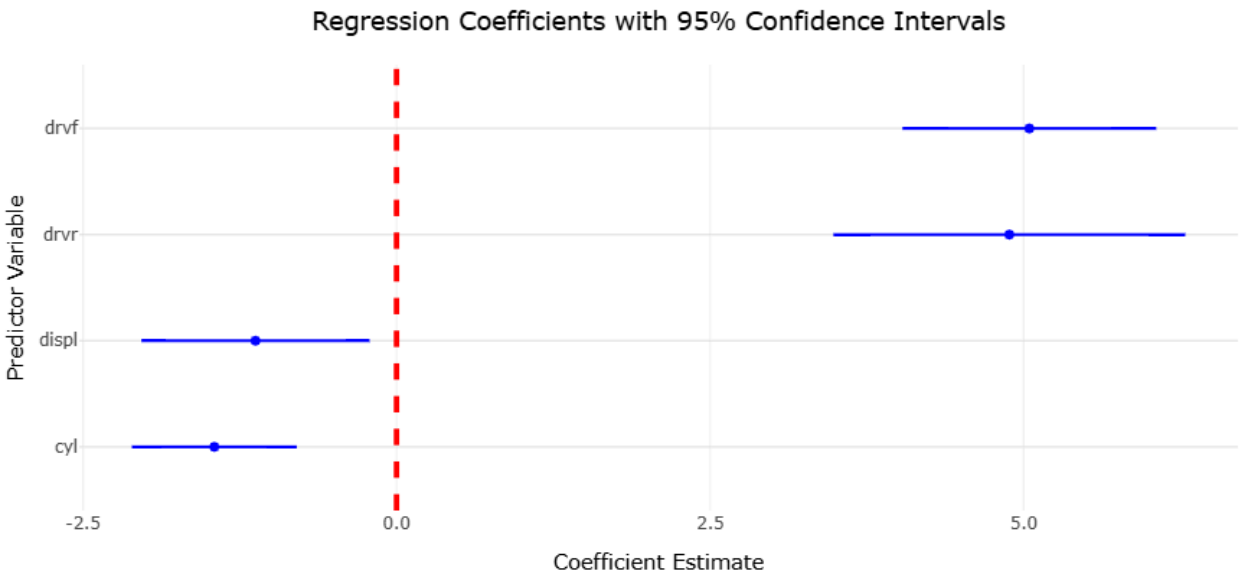
- Prevents Cherry-Picking: Dual-axis plots allow manipulative scaling that can make unrelated trends appear correlated. The separated plots approach prevents this potential misrepresentation.
- Facilitates Direct Comparison: When the plots are aligned vertically with a shared x-axis, the viewer can more accurately compare the timing of changes in both variables without the distraction of conflicting scales.
- Statistical Perspective: The correlation analysis shows that unemployment and personal savings rate have a weak correlation (likely visible in the scatter plot). The dual-axis plot might lead viewers to perceive a stronger relationship than actually exists.

In data visualization best practices, dual-axis charts are generally discouraged because they can be misleading. The patchwork approach provides a clearer, more honest representation of both time series while still allowing for temporal comparisons.

Q3)



Q4)


Regression Coefficients with 95% Confidence Intervals
drvf
drvr
displ
cyl
Predictor Variable
-2.5
0.0
2.5
5.0
Coefficient Estimate