Assignment 2: Basic Techniques

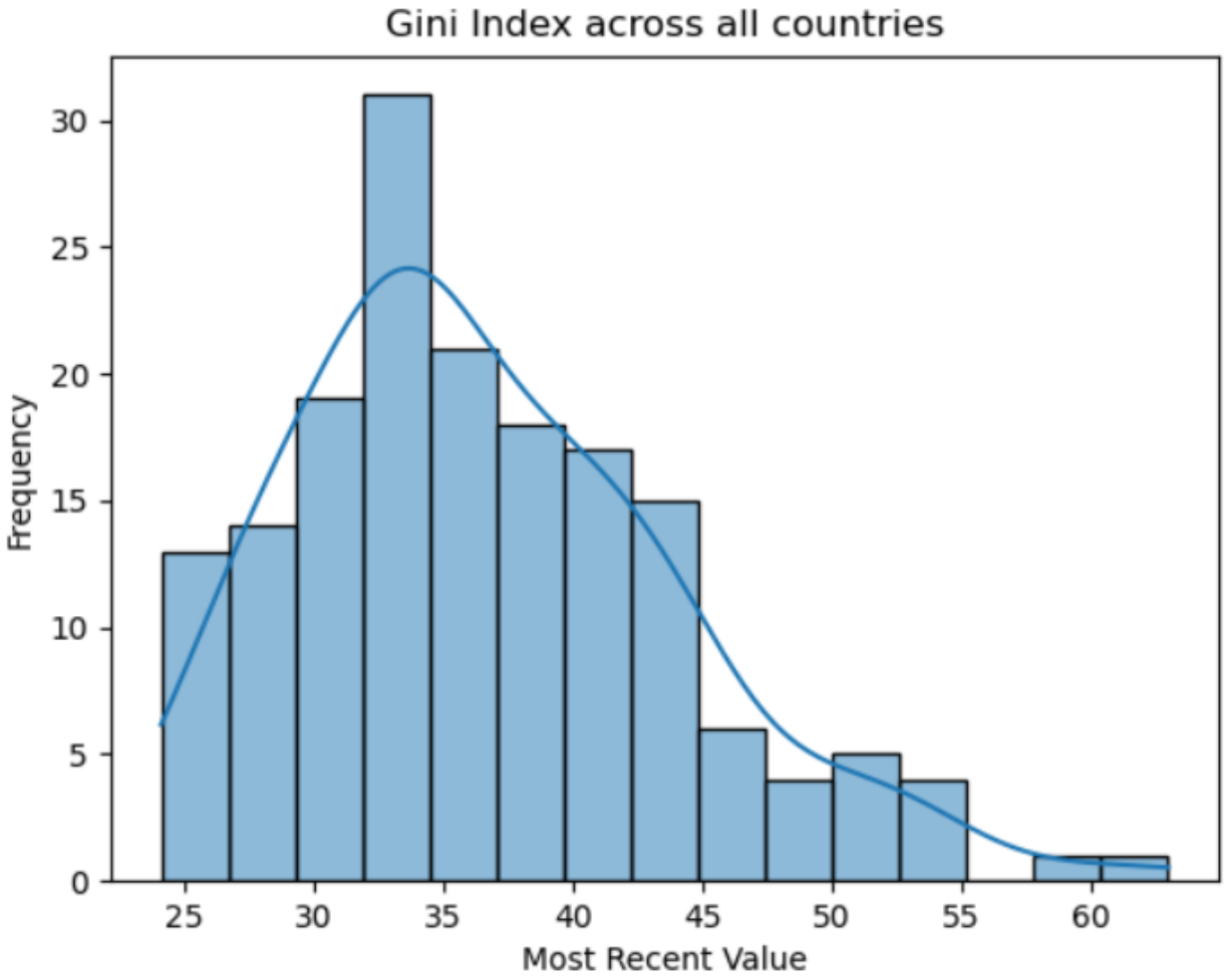
Submission

Note: My chosen SDG in previous submission is **Reduced Inequalities**.

Task 1: Visualizing Distributions

- **Dataset:** Gini index across all countries with available data.
- **Link to the dataset:** https://data.worldbank.org/indicator/SI.POV.GINI
- **Visualization tools:** Python with Matplotlib and Seaborn libraries.

Visualization:



I choose **Gini index** as the variable that is critical for assessing progress toward my chosen SDG **Reduced Inequalities**. Gini index is a number between 0 and 1 or 100, where 0 represents the perfect equality (everyone has the same income). Meanwhile, an index of 1 or 100 represents perfect inequality (one person has all the income, and everyone else has no income). [1]
The above histogram shows the distribution of **'Most Recent Value'** of the Gini index across different nations ranging from approximately 25 to 63. I suspect that this distribution varies across different subgroups of the relevant population. The x-axis represents the Gini index which is binned in intervals (25–30, 30–35, 40–45, …). The y-axis represents the number of countries in each bin. The distribution is right-skewed with a peak around 30-35 and minimum around 55-60. This indicates that most countries have moderate inequality and a few countries have much higher inequality such as South Africa (63) and Namibia (59.1). The histogram shows that the frequencies decrease gradually from 35 to 45, with a sharp drop beyond 45.

The reason I chose the histogram to visualize distribution of Gini index across all countries is because it effectively shows the frequency of countries within specific ranges of inequalities. This hisogram will enable viewers to quickly understand how popular different levels of income inequality are through about 143 countries around the world, which is critical for understanding disparities under SDG **Reduced Inequalities**. Using the '**Most Recent Value**' as x-axis in dataset will provide an overview of current different inequality levels. Moreover, selecting the bin widths is difficult because if bins are too small or too large, it will lead to messy plot or loss of information, respectively. Therefore, the bin width of approximately 5 units (25–30, 30–35, ….) will balance detail and readability, ensuring enough bins to show the variation without overwhelming viewers. Using '**Frequency**' as y-axis will show the number of countries in each inquality level. The scale 0–30 is selected to account for the peak frequency (around 30), ensuring all data is shown without missing any important information. This design ensures the visualization reflects the real-world distribution and draws attention to outliers, prompting further investigation into high-inequality subgroups. In the above histogram, I also include the **Kernel density estimation** (KDE) to smooth the histogram, providing an estimate of probability distribution of data. This enhances the readibility by emphasizing the overall trend and skewness. In addition, I use the light blue color and white background to ensure that viewers can focus on the visualization effectively and the data-ink ratio can be maximized.

The **challenges** I have encountered are that I have to find the dataset related to my chosen SDG. Then I have to find appropriate variable and clean the data. After that, I have to extract these data into the CSV file to analyze and visualize them by programming language Python and its libraries. Visualizing distribution by Seaborn is a bit difficult so the tutorial video on Youtube helps me a lot. I have learned lots of different ways to visualize data through using Seaborn and Matplotlib.

Task 2: Visualizing Time Series

- **Dataset:** Income inequality in Finland (2011–2023) based on Gini coefficient.
- **Link to the dataset:** https://en.wikipedia.org/wiki/List_of_countries_by_income_inequality
- **Visualization tools:** Python with Matplotlib and Seaborn libraries.

Visualization:

I chose **Gini coefficient** as the variable variable that is critical for assessing progress toward my chosen SDG **Reduced Inequalities** and for which I can find temporal data. The temporal data provided (2011–2023) allows us to assess trends in income inequality over time in Finland. The Gini coefficient for Finland shows a slight upward trend from 2011 (25.8) to 2023 (26.6), showing a relatively small increase in income inequality over 13-year period. The value increased by 0.8, which corresponds to about 3.1% (from 25.8 to 26.6). The lowest Gini coefficient value is 25.2 in 2015. The highest values happen in 2022 and 2023 with the same value 26.6, which is a peak in income inequality from 2011 to 2023. There are small fluctuations from 2011 to 2017 (around 25.2–25.6), followed by a gradual increase. From 2018, the Gini coefficient increased from about 25.9 to 26.6 in 2023, with a slight dip in 2021 (25.7), suggesting a potential worsening of income inequality in the most recent years. In general, the Gini coefficient for Finland remains relatively low compared to global standards (30–35 in many other countries). This indicates that Finland maintains a relatively equitable income distribution. However, the slight upward trend suggests that the level of inequality is increasing. The reasons might be the change of policies and the outcome of Covid-19 pandemic.

The reason I chose the dot plot to visualize time series of Gini coefficient for Finland (2011–2023) is because I need to emphasize discrete data points over time. I do not use the line graphs, which is typically the best choice to display time series, because intermediate values are not guaranteed

to be close to linear interpolation. Using dots highlights individual data points without implying a continuous trend, which is appropriate in this scenario. A cluttered design could distract from the key insight of the visualization. For assessing progress toward reducing inequalities, it is critical to see how inequality levels fluctuate annually. Dots allow viewers to focus on specific years (the dip in 2015 or the peak in 2022–2023) and assess whether progress is consistent or fluctuated, which is more relevant than a trend line. I use red dots and white background to make viewers focus on the data points without overwhelming them with additional visual elements and maximize the data-ink ratio. Labeling the x-axis with years (2011–2023) and the y-axis with the Gini coefficient scale provides immediate context. I rotate x-axis labels for angle 45° to make better readability between each consecutive year. The design aligns with the lecture's emphasis on time-series visualization, using techniques like clear labeling to support analysis, while the focus on dots reflects a good choice to highlight individual data points for the chosen SDG.

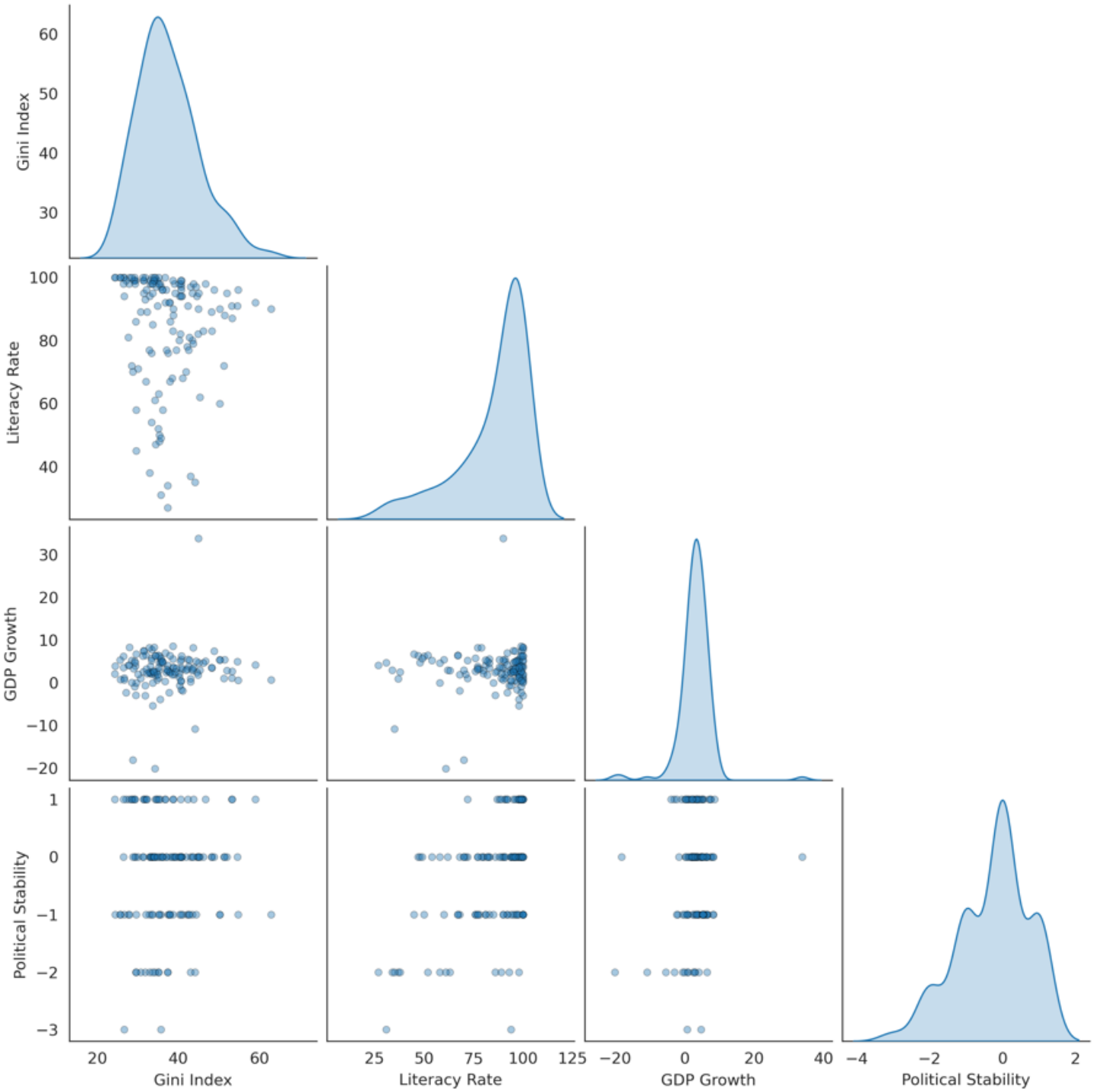
The **challenges** I have encountered are that I am confused between line graphs or dot plots. I do not know which choice is better. However, fortunately, I find that dot plots should be used when measurements are at irregular intervals and intermediate values are not guaranteed close to be linear interpolation. Therefore, I finally chose dot plot as my design choice.

Task 3: Visualizing High-Dimensional Data

- **Dataset:**
  1. GDP growth (annual %)
  2. Gini index
  3. Literacy rate, adult total (% of people ages 15 and above)
  4. Political Stability and Absence of Violence/Terrorism
- **Link to the dataset, respectively:**
  1. https://data.worldbank.org/indicator/NY.GDP.MKTP.KD.ZG
  2. https://data.worldbank.org/indicator/SI.POV.GINI
  3. https://data.worldbank.org/indicator/SE.ADT.LITR.ZS
  4. https://data.worldbank.org/indicator/PV.EST
- **Visualization tools:** Python with Matplotlib and Seaborn libraries.

Challenges: The task requires to identify four variables so I have to find four different datasets as above. Then I have to merge these datasets into one CSV file to analyze and visualize the information. This process is a bit complicated because of some configured errors. Fortunately, these datasets are from the same source, so I do not waste so much time to find and analyze them. I have to follow the ChatGPT's instructions to merge these datasets into one CSV file and help me how to plot the high-dimensional data using Matplotlib and Seaborn libraries because I have never written code related to this type of plot before. After that, I updated the code to improve the visualization based on the lecture and recording from the teachers. I also have to follow the tutorial video on Youtube to learn some new knowledge related to how to visualize high-dimensional data, which is very helpful for doing this task. In general, doing this task is very challenging but helps me to learn lots of things related to high-dimensional data.

Visualization:



I chose **small multiples** to visualize these four variables. The visualization consists of a 4x4 matrix where each row and each column corresponds to one of four variables: Gini Index, Literacy Rate, GDP Growth, and Political Stability. The diagonal elements show the distributions of each variable and the off-diagonal elements display scatter plots to show the pairwise relationships between the variables.

For **Gini Index and Literacy Rate** (row 2, column 1), the scatter plot shows a cluster of points where higher literacy rates (closer to 100%) correspond to a wide range of Gini Index values (from 20 to 40). However, for lower literacy rates (below 80%), the Gini Index tends to be higher (around 40–60). This shows a negative relationship: countries with lower literacy rates tend to have higher income inequality or higher Gini Index. However, high literacy rates do not guarantee low inequality, as the Gini Index varies widely even at high literacy levels.

For **Gini Index and GDP Growth** (row 3, column 1), the scatter plot shows that the data points are loosely clustered. The trend is not strong. The range of data points concentrates from about -10% to 10%. There might be a slight negative correlation: as Gini Index increases, GDP growth tends to decline in some cases.

For **Gini Index and Political Stability** (row 4, column 1), the scatter plot shows that higher Gini Index values tend to correspond to lower political stability (about –1). In contrast, lower Gini Index values are associated with higher political stability (about 0). This shows that there might be a slight negative relationship: higher income inequality is associated with lower political stability, suggesting that inequality may result in political instability.

For **Literacy Rate and GDP Growth** (row 3, column 2), there is a weak positive relationship: countries with higher literacy rates are more likely to experience positive GDP growth, but the relationship is not strong, as high literacy rates still correspond to a broad range of growth outcomes.

For **Literacy Rate and Political Stability** (row 4, column 2), the scatter plot shows that higher literacy rates (closer to 100%) are associated with higher political stability (closer to 1). Lower literacy rates (below 50%) tend to correspond to lower stability (closer to –2). There is a positive relationship: higher literacy rates are associated with greater political stability.

For **GDP Growth and Political Stability** (row 4, column 3), there is a weak positive relationship: political stability seems to support economic growth, while instability is associated with economic decline.

The reason I chose **small multiples** to visualize high-dimensional data is because it is the best option to explore the pairwise relationships between variables. I arrange the small multiples into a lower triangular matrix to avoid repeating the pairwise relationships. I also include histogram

on the diagonal to show each distribution of each variable. Understanding the shape of each variable's distribution helps viewers interpret the scatter plots more effectively. For example, knowing that the Literacy Rate is heavily skewed toward 100% explains why most data points in the scatter plots involving Literacy Rate are clustered at the higher end of the axis. The axes for each variable are consistent across all plots so that comparisons are fair and the viewer can easily interpret the scale. I also try to reduce clutter by reducing the marker size and making markers transparent (which helps visualize overlapping data points). I uses a simple color palette and white background, which keeps the focus on the data rather than distracting with unnecessary colors and maximizes the data-ink ratio.

References:

[1] *Wikipedia* . **List of countries by income inequality**. Available at: https://en.wikipedia.org/wiki/List_of_countries_by_income_inequality