**Statistics Analysis for Population Dataset:**

**Step 1: Summarize the Dataset**

1. Data Preprocessing:
   * Load the population dataset and address any missing data, represented as '-' or NaN
   * Convert data types as needed (e.g., 'Year' to integer, 'VALUE' to floating-point).
2. Descriptive Statistics:
   * Calculate key descriptive statistics for the 'VALUE' column:
     + **Mean**: The average population across years.
     + **Median**: The middle value, providing insights into central tendencies.
     + **Standard Deviation**: Measure of population data variability.
   * Justify the relevance of these statistics.
3. Data Visualization:
   * Create at least three relevant data visualizations:
     + **Line Plot**: Depicting population changes over time ('Year' vs. 'VALUE').
     + **Age Distribution**: Visualizing population distribution by age.

Example using **histogram** or **bar plot.**

# Filter the data for a specific year and 'Both sexes' for sex

year\_to\_visualize = 2023

sex\_to\_visualize = 'Both sexes'

filtered\_data = population\_df[(population\_df['Year'] == year\_to\_visualize) & (population\_df['Sex'] == sex\_to\_visualize) & (population\_df['Single Year of Age'] != 'All ages')]

# Plot the population distribution by age

age\_groups = filtered\_data['Single Year of Age'].unique()

xtick\_positions = [0, len(age\_groups) - 1]

for i in range(5, len(age\_groups) - 1, 5):

xtick\_positions.append(i)

plt.figure(figsize=(10, 6))

#plt.bar(filtered\_data['Single Year of Age'], filtered\_data['VALUE'])

plt.bar(range(len(age\_groups)), filtered\_data['VALUE'])

plt.xlabel('Age')

plt.ylabel('Population')

plt.title(f'Population Distribution by Age in {year\_to\_visualize} ({sex\_to\_visualize})')

#plt.xticks(filtered\_data['Single Year of Age'][::5], rotation=90)

plt.xticks(xtick\_positions, [age\_groups[i] for i in xtick\_positions], rotation=90)

plt.tight\_layout()

plt.show()

**A graph of a number of people

Description automatically generated**

* + - **Population by Sex**: Showcasing the distribution of 'VALUE' by sex.

Example:

years\_to\_visualize = range(1926, 2023, 5)

sex\_to\_visualize = 'Both sexes'

male\_populations = []

female\_populations = []

for year in years\_to\_visualize:

# Filter the data for the specific year and 'Both sexes' for sex

year\_data = population\_df[(population\_df['Year'] == year) & (population\_df['Sex'] == sex\_to\_visualize)]

# Separate the data into male and female populations

male\_data = year\_data[year\_data['Single Year of Age'] != 'All ages']

male\_population = male\_data['VALUE'].sum()

male\_populations.append(male\_population)

female\_data = year\_data[year\_data['Single Year of Age'] != 'All ages']

female\_population = female\_data['VALUE'].sum()

female\_populations.append(female\_population)

plt.figure(figsize=(12, 6))

plt.bar(years\_to\_visualize, male\_populations, label='Male')

plt.bar(years\_to\_visualize, female\_populations, label='Female', bottom=male\_populations)

plt.xlabel('Year')

plt.ylabel('Population')

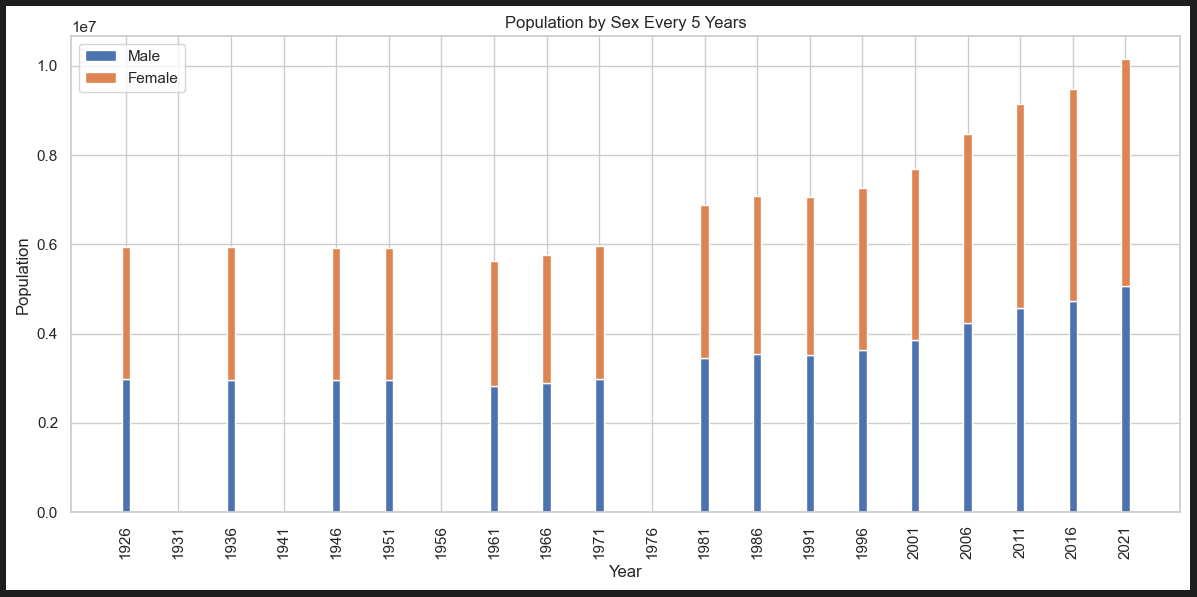
plt.title('Population by Sex Every 5 Years')

plt.xticks(years\_to\_visualize, rotation=90)

plt.legend()

plt.tight\_layout()

plt.show()



* + Critically analyze the findings from these visualizations.

***Step 2: Discrete Distributions***

4. Select Discrete Distributions:

* Choose between Binomial and Poisson distributions based on the dataset's characteristics.
* Justify your selection based on the nature of the 'VALUE' data.

1. Application of Discrete Distributions:
   * Apply the selected distribution(s) to model the 'VALUE' data.
   * Use Python to perform this modeling.
2. Large Sample Behavior:
   * Explore and visualize how the chosen distribution(s) behave with large samples.
   * Explain how this behavior relates to the population dataset.

*Step 3: Normal Distribution (0-20)* 7. Applicability of Normal Distribution:

* Investigate whether the Normal distribution is suitable for representing the 'VALUE' data.
* Explain the rationale behind considering a Normal distribution.

1. Fit Normal Distribution:
   * Apply Python to fit a Normal distribution to the 'VALUE' data.
   * Visualize the distribution's fit to actual population data.

*Step 4: Explain Importance (0-15)* 9. Explain the Significance:

* Elaborate on the importance of selecting specific distributions.
* Provide reasoning behind the choice of variables in the analysis.

1. Evaluation of Variables:
   * Assess whether 'VALUE' can be effectively modeled using Normal distribution characteristics.
   * Discuss any potential limitations or advantages of this modeling approach.

**Statistics Analysis for Migration Dataset:**

*Step 1: Summarize the Dataset (0-35)*

1. Data Preprocessing:
   * Load the migration dataset and handle any missing data represented as '-'.
   * Ensure proper data types, such as 'Year' to integer and 'VALUE' to floating-point.
2. Descriptive Statistics:
   * Calculate descriptive statistics for the 'VALUE' column, including mean, median, and standard deviation.
   * Explain the relevance of these statistics in the context of migration.
3. Data Visualization:
   * Create relevant data visualizations:
     + Time Series Plot: Illustrate net migration trends over time ('Year' vs. 'VALUE').
     + Migration by Country: Visualize migration patterns between countries.
   * Analyze the insights obtained from these visualizations.

*Step 2: Discrete Distributions (0-30)* 4. Select Discrete Distributions:

* Decide if Binomial and/or Poisson distributions are appropriate for describing aspects of migration data.
* Provide reasoning for your choice based on specific data characteristics.

1. Application of Discrete Distributions:
   * Apply the chosen distribution(s) to model the 'VALUE' data.
   * Implement this modeling using Python.
2. Large Sample Behavior:
   * Examine how the selected distribution(s) behave with larger sample sizes.
   * Explain how this behavior is relevant to migration data.

*Step 3: Normal Distribution (0-20)* 7. Applicability of Normal Distribution:

* Explore whether a Normal distribution can offer insights into the 'VALUE' data.
* Describe the motivation behind considering a Normal distribution.

1. Fit Normal Distribution:
   * Employ Python to fit a Normal distribution to the 'VALUE' data.
   * Assess the fit of this distribution to the actual migration data.

*Step 4: Explain Importance (0-15)* 9. Explain the Significance:

* Discuss the importance of the selected distributions in your analysis.
* Justify your choice of variables and evaluate their suitability for modeling with Normal distribution characteristics.

This comprehensive plan provides detailed instructions for each step of the statistical analysis of both the population and migration datasets, ensuring a thorough exploration of the data and rigorous application of statistical techniques.