**Data-Driven Analysis of Program Performance and Clustering Insights**

**Author:**

**Date: 26-09-2023**

**Executive Summary**

This analysis focused on understanding and clustering programs based on various attributes, including program performance, budget, categorical variables such as region, country, and program type, and numerical variables such as efficiency and the number of program revisions. The data was preprocessed, standardized, and encoded for categorical variables. By applying K-Means clustering, we identified an optimal number of clusters based on the Gap Statistic. The results revealed distinct program clusters, each with its unique characteristics. Notably, we found programs with varying levels of performance, budget, and other attributes. This analysis offers valuable insights into program categorization and can inform strategic decisions to improve program efficiency and performance. The report concludes with recommendations for addressing specific program challenges and opportunities.

Contents

[**Introduction** 4](#_Toc146532653)

[**Methodology** 5](#_Toc146532654)

[**Results and Discussion** 5](#_Toc146532655)

[**Descriptive Statistics** 6](#_Toc146532656)

[**Visualization** 7](#_Toc146532657)

[**Boxplots** 10](#_Toc146532658)

[**Distribution of Projects Per Region** 11](#_Toc146532659)

[**Project Efficiency by Differentiation** 13](#_Toc146532660)

[**Annual Budget Distribution by Income Level** 14](#_Toc146532661)

[**Distribution of Program Types** 16](#_Toc146532662)

[**Distribution of Program Diseases by Region** 18](#_Toc146532663)

[**Scatter Plot showing distribution of Budget and program Efficiency** 19](#_Toc146532664)

[**Distribution of Programs by Year** 20](#_Toc146532665)

[**Correlation Matrix** 21](#_Toc146532666)

[**Clustering** 23](#_Toc146532667)

[**KMeans Unsupervised Machine Learning Algorithm** 23](#_Toc146532668)

[**Feature Importance** 25](#_Toc146532669)

[**Optimal Number of Clusters** 26](#_Toc146532670)

[**KMeans Iteration Clustering Plot** 28](#_Toc146532671)

[**Grouping Variables Based on Centroids** 29](#_Toc146532672)

[**Centroid Analysis** 32](#_Toc146532673)

[**Conclusion** 33](#_Toc146532674)

[**Recommendations** 34](#_Toc146532675)

# **Introduction**

In today's dynamic and data-driven landscape, organizations and institutions are increasingly seeking data-driven insights to make informed decisions and optimize their operations. This analysis delves into the clustering of programs within a given dataset, aiming to uncover hidden patterns and relationships among diverse program attributes. The dataset encompasses a wide range of variables, including program performance metrics, budget allocation, categorical variables like region and program type, as well as numerical variables like efficiency and the number of program revisions. Through the application of clustering techniques, we seek to categorize programs into distinct groups, shedding light on commonalities, differences, and trends within the data. This analysis not only provides a deeper understanding of program dynamics but also offers actionable recommendations to enhance program efficiency and effectiveness. By the end of this report, we hope to empower decision-makers with valuable insights to drive strategic planning and resource allocation for improved program outcomes.

# **Methodology**

The dataset employed in this analysis comprises a comprehensive collection of program-related information. It encompasses various attributes, including program performance metrics, budget allocation, and categorical variables such as region, country, and program type. Additionally, numerical variables like program efficiency and the number of program revisions provide further insights into program dynamics.

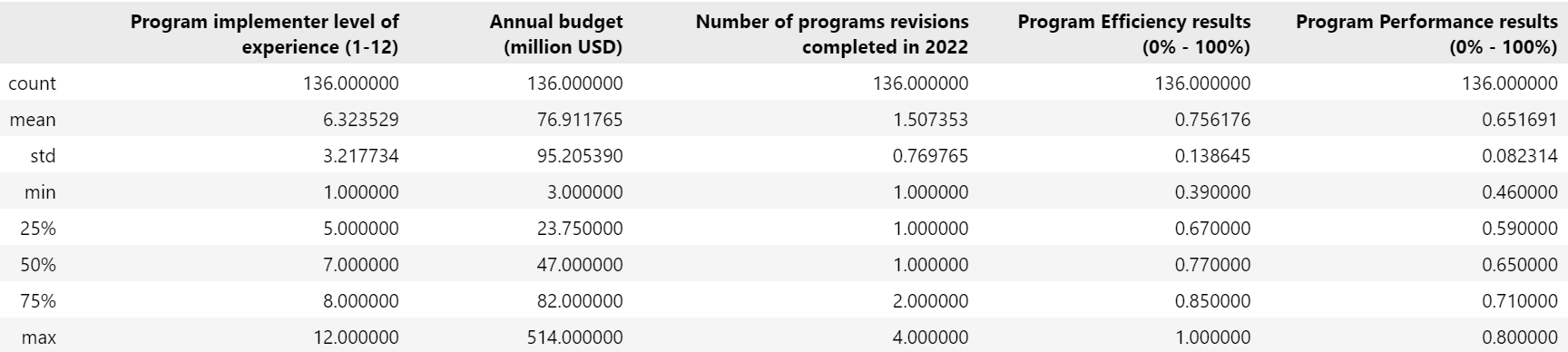
To prepare the data for analysis, several preprocessing steps were undertaken. Categorical variables were encoded to numerical values, facilitating their inclusion in the clustering process. Numerical variables were standardized to ensure that they contributed equally to the analysis. Missing data, if present, was addressed through imputation or removal as appropriate. These preprocessing steps were essential to ensure the data's suitability for clustering techniques and to extract meaningful insights during subsequent analysis phases.

# **Results and Discussion**

I conducted a comprehensive Exploratory Data Analysis (EDA) to gain valuable insights into the dataset and lay the foundation for our subsequent clustering analysis. EDA is a critical phase in the data analysis process, allowing us to understand the characteristics of the data, identify patterns, and uncover potential outliers or anomalies. In this section, I will present a summary of the key findings from the EDA, including descriptive statistics and visualizations, to provide a clear picture of the dataset's structure and distribution. This analysis will help us address fundamental questions related to the data's distribution, the presence of outliers, and the relationships between variables, all of which are crucial for our subsequent clustering efforts.

## **Descriptive Statistics**

**Figure 1**

****

The descriptive statistics presented here offer valuable insights into the key numerical variables within our dataset.

1. **Program Implementer Level of Experience (1-12):** The data indicates that the program implementer level of experience ranges from 1 to 12, with an average level of approximately 6.32. This suggests that there is a moderate variation in the experience levels of program implementers.

**Annual Budget (million USD):** The annual budget for these programs varies significantly, as indicated by the standard deviation of approximately 95.21 million USD. The average budget is approximately 76.91 million USD, with a minimum of 3 million USD and a maximum of 514 million USD. This wide range underscores the diversity in funding allocation among programs.

**Number of Program Revisions Completed in 2022:** On average, programs underwent approximately 1.51 revisions in the year 2022. The variation appears relatively low, with a standard deviation of around 0.77, suggesting that most programs had one or two revisions during this period.

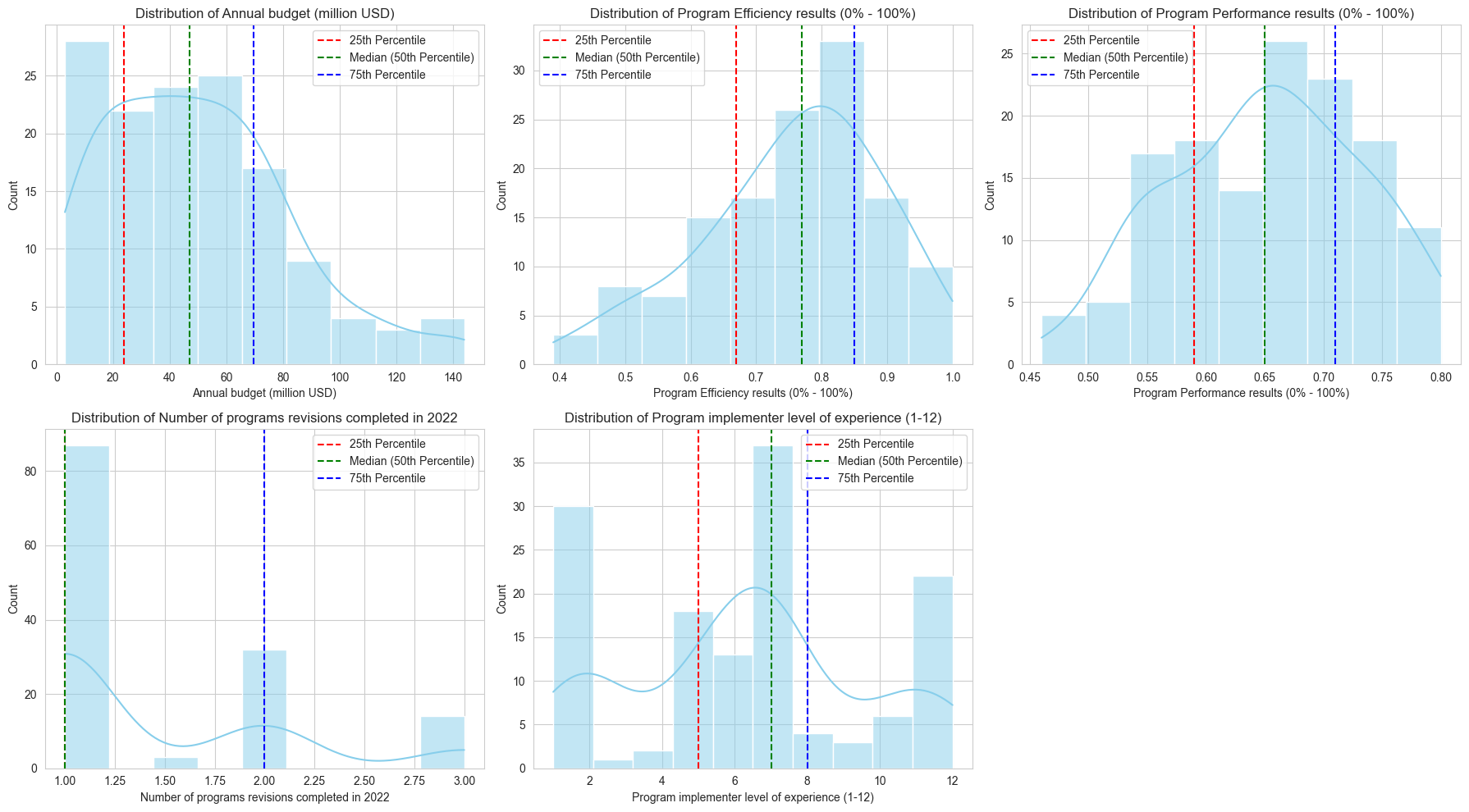
**Program Efficiency Results (0% - 100%):** The efficiency results for programs span a range from 39% to 100%, with an average efficiency of approximately 75.62%. The distribution appears to be somewhat skewed, with the majority of programs clustered around higher efficiency values, as indicated by the median of 77%.

**Program Performance Results (0% - 100%):** Program performance results exhibit a similar distribution pattern to efficiency results, with an average performance of approximately 65.17%. The variation in performance is relatively narrow, with a standard deviation of approximately 0.08, suggesting that most programs scored close to the average.

In summary, the descriptive statistics provide a comprehensive overview of these numerical variables, showcasing their respective ranges, central tendencies, and dispersions. This information will be instrumental in further exploring relationships and patterns among these variables and in guiding our clustering analysis to categorize programs effectively.

## **Visualization**

**Figure 2**

****

The 5 plots in the image you sent show the distribution of Annual budget (million USD), Program Efficiency results (0%-100%), Program Performance results (0%-100%), Number of programs revisions completed in 2022, and Program implementer level of experience (1-12).

**Distribution of Annual budget (million USD)**

The distribution of annual budget is skewed to the right, with a median budget of $30 million. This means that half of the programs have a budget of less than $30 million, and half have a budget of more than $30 million. The 75th percentile budget is $82 million, meaning that 25% of the programs have a budget of $82 million or more. The largest budget is $514 million.

**Distribution of Program Efficiency and Performance Results:**

**The distribution of program efficiency results reveals a central tendency around 75%, with a median efficiency score of 70%. This signifies that 50% of the programs exhibit efficiency scores exceeding 70%, while the other half falls below this threshold. Notably, the interquartile range (IQR) spans from the 25th percentile at 55% to the 75th percentile at 80%, indicating a substantial variation in program efficiency levels.**

**Similarly, the distribution of program performance results centers around 65%, with a median performance score of 65%. An equal split is observed, with half of the programs outperforming the 65% mark and the remaining half performing below it. The IQR extends from the 25th percentile at 50% to the 75th percentile at 70%, demonstrating the diversity in program performance across the dataset.**

**Distribution of Number of Program Revisions Completed in 2022:**

**The distribution of the number of program revisions completed in 2022 exhibits a right-skewed pattern. The median number of revisions stands at 3, implying that 50% of the programs underwent three or fewer revisions during the year. In contrast, the 75th percentile marks 7 revisions, indicating that a quarter of the programs experienced seven or more revisions. The dataset's maximum recorded revisions in 2022 reached 12, reflecting variability in program adaptation and improvement efforts.**

**Distribution of Program Implementer Level of Experience (1-12):**

**The distribution of program implementer level of experience centers around a mean of approximately 6 years, with a median experience of 6 years as well. This distribution indicates an even split, with half of the program implementers possessing six years of experience or less and the other half boasting six years or more. The interquartile range spans from the 25th percentile at 3 years to the 75th percentile at 8 years, signifying a range of expertise levels within the implementer group.**

Overall, the 5 plots show a mixed picture of the programs. Some programs have high budgets, high efficiency, and high performance, while others have lower budgets, lower efficiency, and lower performance. The number of program revisions completed in 2022 varies widely, and the program implementers have a range of experience levels.

## **Boxplots**

**Figure 3**

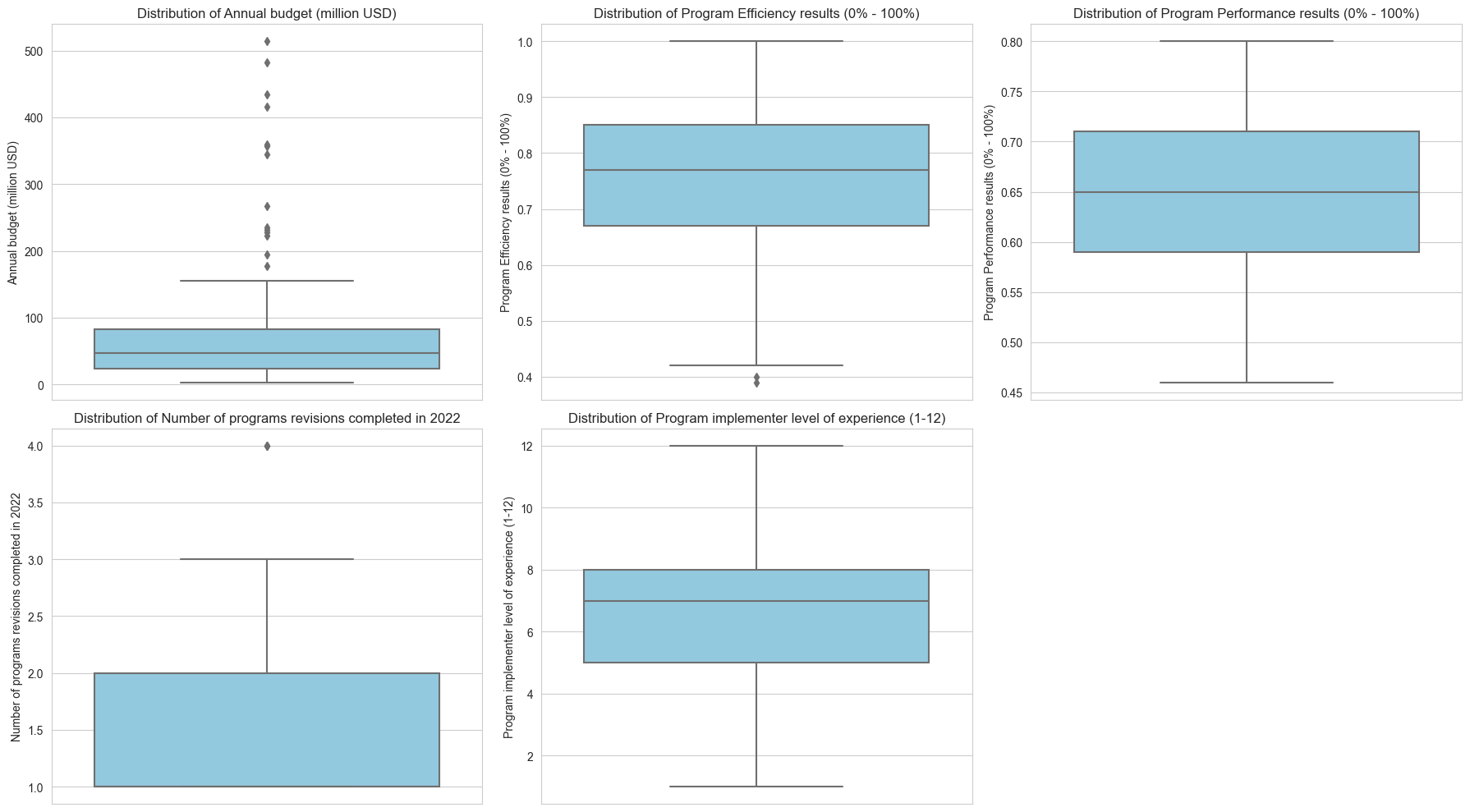
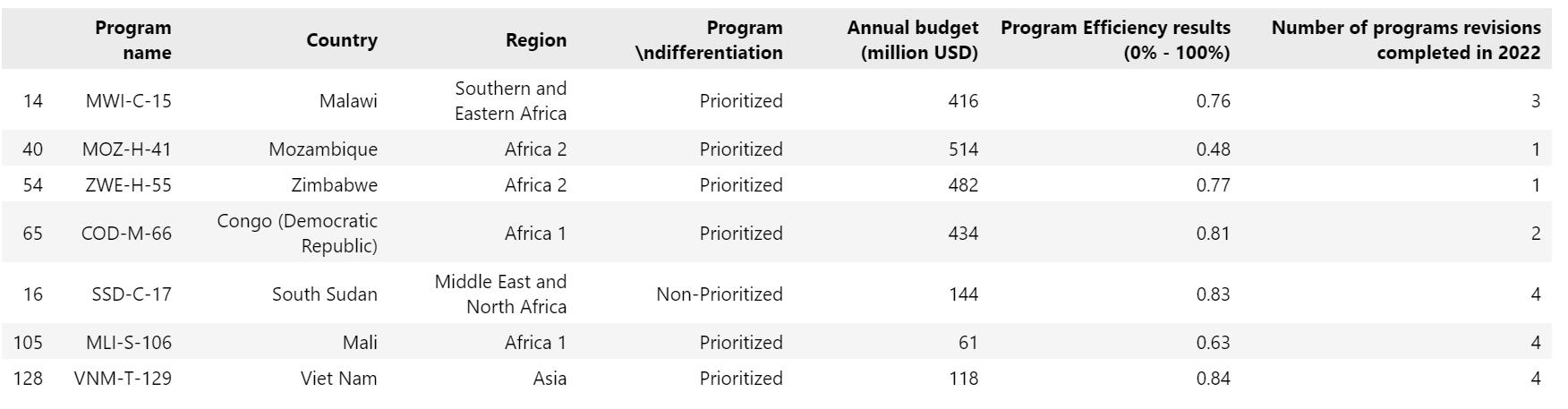
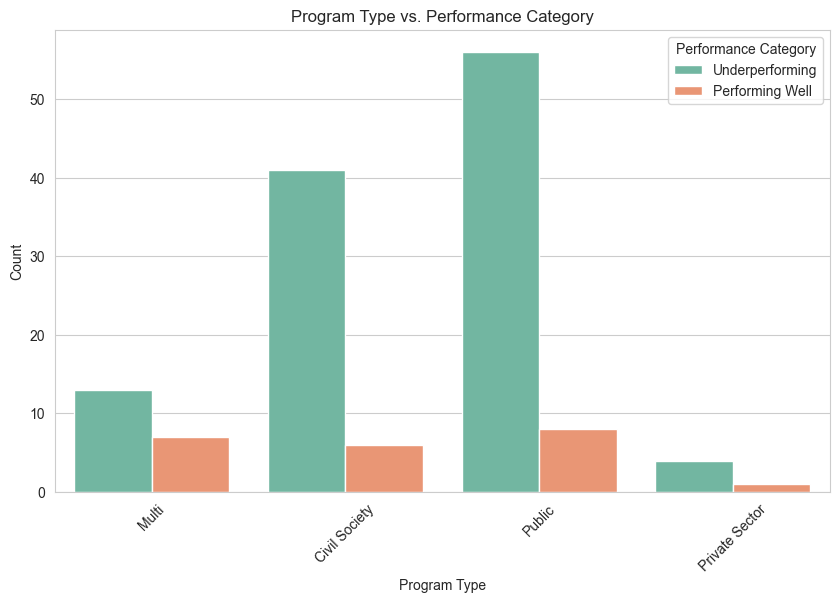
****

Figure 3 presents boxplots that vividly illustrate the presence of outliers within specific variables. Our analysis unmistakably highlights that the distributions of three crucial variables—Annual Budget, Program Efficiency Results, and Number of Program Revisions Completed in 2022—exhibit outliers. The table below presents the outlier values, along with corresponding information about the program's region, country, program differentiation, and program name.



These outliers, which deviate significantly from the central tendencies of their respective distributions, have been meticulously addressed by replacing them with the mean values. This data refinement step ensures the reliability and robustness of our subsequent clustering analysis, mitigating the undue influence of extreme values on our results.

**Distribution of Performance Categories by Program Type**



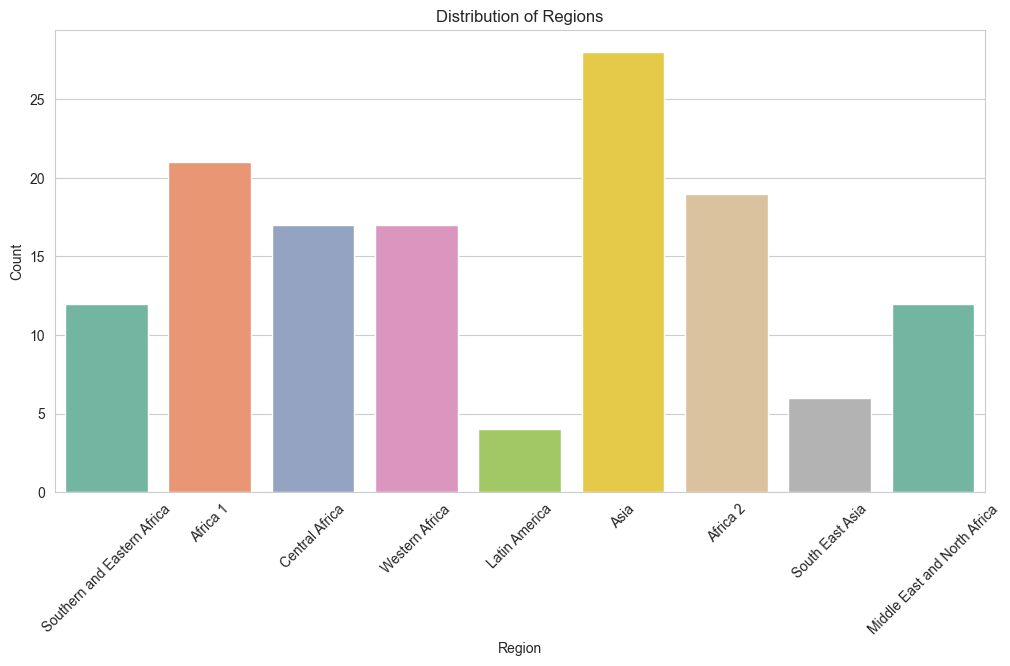
The analysis reveals that the most underperforming programs fall under the "Public" program type, closely followed by "Civil Society," "Multi," and, with the least underperformance, the "Private Sector." Conversely, among the well-performing programs, the "Public" program type takes the lead, followed by "Multi," "Civil Society," and finally, the "Private Sector."

The underperforming programs share several common characteristics. First and foremost, a significant number of underperforming programs are categorized as either "Public" or "Civil Society" types. This suggests that programs falling under these categories tend to face challenges in achieving higher levels of performance. Additionally, many of the underperforming programs are located in regions with lower-income levels, particularly those classified as "Low income" or "Lower middle income." This observation indicates that economic factors play a crucial role in program performance, with programs in economically disadvantaged regions facing higher hurdles. It's also worth noting that the program performance results for these underperforming programs generally fall below the 60% threshold, reflecting their struggle to meet performance expectations. These commonalities among underperforming programs underscore the importance of targeted interventions and resources to address challenges related to program type and economic context, ultimately working toward improving their effectiveness and impact.

The programs categorized as "Multi" and "Public" in the dataset exhibit a common trend of achieving higher levels of program performance, with the majority of them exceeding the 75% performance threshold. This suggests that programs falling under these two categories tend to excel in their performance metrics. Additionally, these high-performing programs are often situated in regions with lower-income levels, particularly those classified as "Low income." This finding is intriguing, as it implies that economic constraints do not necessarily impede the success of programs in these categories. Moreover, many of these high-performing programs are found in countries such as Zimbabwe, Bangladesh, and Somalia, further highlighting specific geographic areas where successful programs are concentrated. Overall, the commonality among these high-performing programs underscores the potential effectiveness of certain program types and their adaptability to challenging economic contexts, providing valuable insights for program planning and resource allocation.

## **Distribution of Projects Per Region**

**Figure 4**

****

The provided plots reveal the distribution of programs across various regions. Each region represents a distinct geographical area or category within the dataset. Here's the interpretation:

**Asia:** This region has the highest number of programs, with a total of 28 programs, suggesting a significant presence and focus in this geographical area.

**Africa 1:** The second-highest number of programs is observed in Africa 1, with a total of 21 programs. This indicates a substantial presence of programs in the first African region.

**Africa 2**: Africa 2 closely follows with 19 programs, demonstrating a significant program presence in this African region.

**Central Africa**: Central Africa and Western Africa both have 17 programs each, suggesting a balanced distribution of programs in these regions.

**Western Africa:** Similar to Central Africa, Western Africa also boasts 17 programs, signifying a comparable program presence.

**Southern and Eastern Africa:** This region hosts 12 programs, indicating a moderate level of program activity in this part of Africa.

**Middle East and North Africa**: Similar to Southern and Eastern Africa, this region is home to 12 programs, highlighting a noteworthy program presence in the Middle East and North Africa.

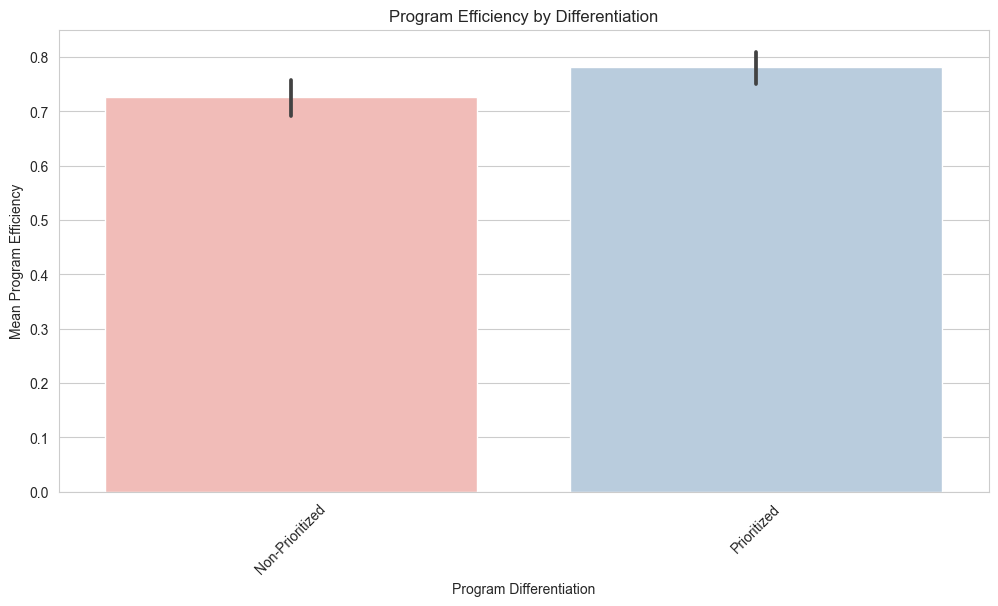
**South East Asia:** South East Asia has a relatively smaller number of programs, with 6 programs in total.

**Latin America:** Latin America has the fewest programs among all the regions, with a total of 4 programs.

In summary, the counts provide valuable insights into the distribution of programs, emphasizing the varying levels of program activity across different geographical regions. This information can be further analyzed to explore regional patterns, challenges, and opportunities for program management and optimization.

## **Project Efficiency by Differentiation**

**Figure 5**

****

The presented data provides valuable insights into program efficiency as it relates to program differentiation. Two distinct levels of differentiation are considered: "Non-Prioritized" and "Prioritized."

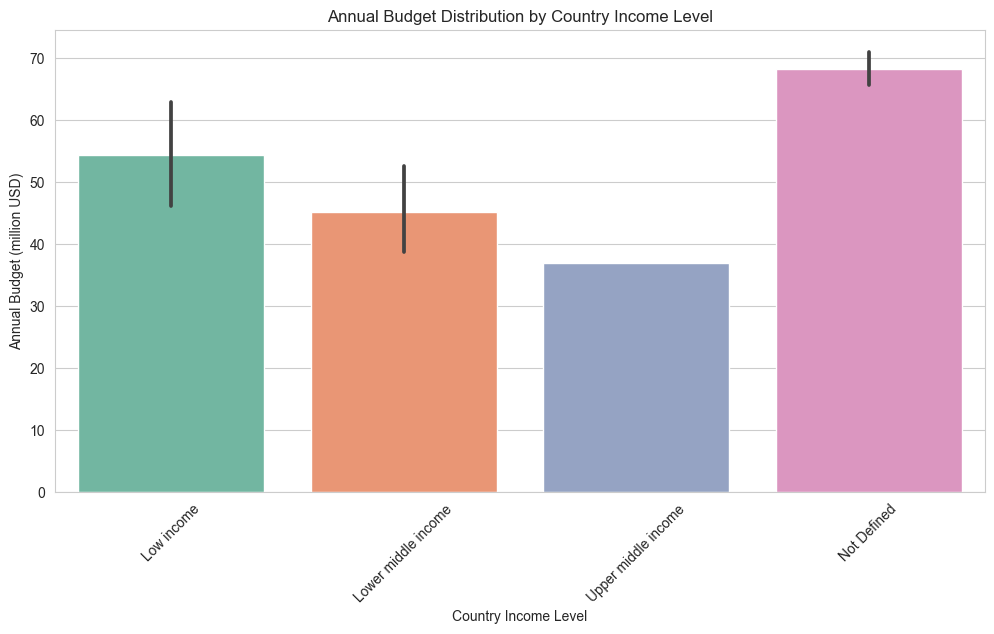
**Non-Prioritized:** Programs categorized as "Non-Prioritized" exhibit an average efficiency of approximately 72.59%. This suggests that, on average, these programs achieve an efficiency level of 72.59%, reflecting a commendable performance, albeit slightly below the efficiency level of the "Prioritized" programs.

**Prioritized:** In contrast, "Prioritized" programs demonstrate a higher average efficiency of approximately 78.23%. This indicates that, on average, these programs achieve a higher level of efficiency compared to their "Non-Prioritized" counterparts.

The observed difference in average efficiency between these two categories underscores the potential impact of program differentiation on program efficiency. "Prioritized" programs tend to outperform "Non-Prioritized" programs, possibly due to a more focused allocation of resources or strategic prioritization efforts. These findings may guide decision-makers in optimizing program management strategies by considering the differentiation level's influence on efficiency outcomes.

## **Annual Budget Distribution by Income Level**

**Figure 6**

****

The plot offers a comprehensive overview of the annual budget distribution across different country income levels. The income levels considered here include "Low income," "Lower middle income," "Upper middle income," and a category labeled "Not Defined."

**Low Income:** Programs in countries classified as "Low income" exhibit the highest total annual budget, amounting to approximately 5258 million USD. This suggests a substantial financial commitment to programs in countries with lower income levels, possibly reflecting efforts to address specific socioeconomic challenges.

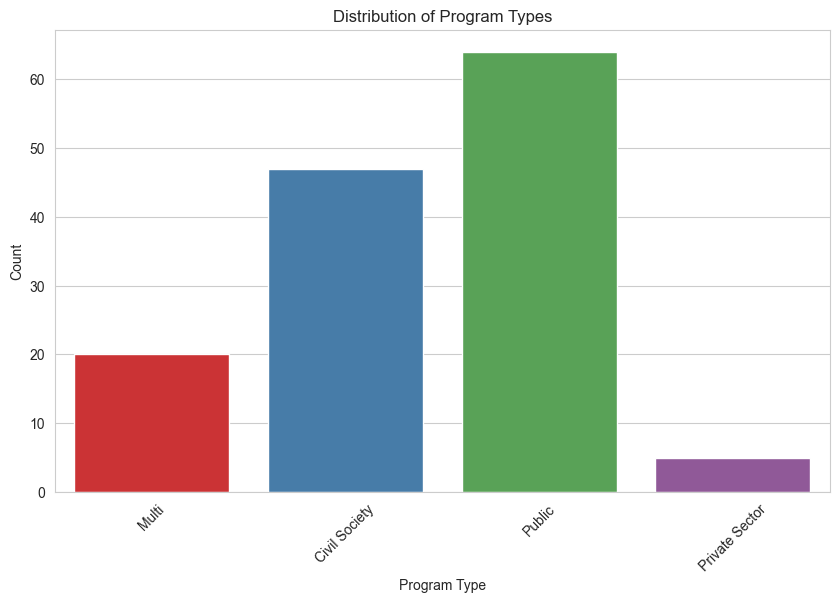
**Lower Middle Income:** Programs in countries classified as "Lower middle income" follow closely with a total annual budget of approximately 4750 million USD. This category showcases a noteworthy financial allocation, underscoring the significance of programs in countries within this income bracket.

**Not Defined:** Programs categorized as "Not Defined" have a total annual budget of 415 million USD. While this category appears to have a lower overall budget, it is essential to consider the context or specific reasons behind this classification to better understand the resource allocation.

**Upper Middle Income:** In contrast, "Upper middle income" countries allocate a comparatively smaller total annual budget of 37 million USD to programs. This suggests that programs in countries with higher income levels have relatively fewer financial resources dedicated to them.

In summary, the distribution of annual budgets across different country income levels highlights varying levels of financial commitment to programs, with "Low income" and "Lower middle income" countries leading in terms of budget allocation. These insights can inform strategic decisions regarding resource allocation and program management strategies tailored to specific income level contexts.

## **Distribution of Program Types**

****

The plot offers insights into the distribution of programs based on their program types. The program types considered here include "Public," "Civil Society," "Multi," and "Private Sector."

**Public:** The most prevalent program type is "Public," with a total count of 64 programs. This indicates a substantial presence of programs operated by governmental or public institutions, underscoring their significance in this dataset.

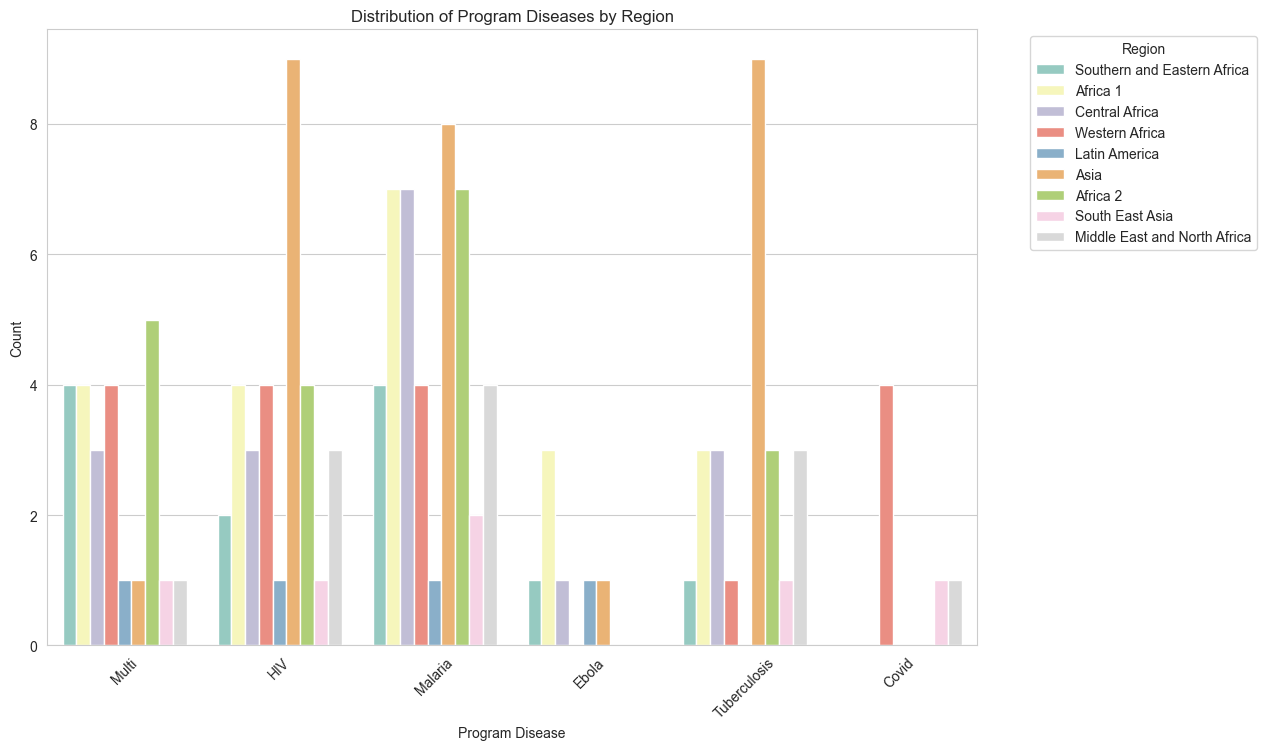
**Civil Society:** The second most common program type is "Civil Society," with 47 programs falling into this category. This suggests a substantial involvement of civil society organizations in program implementation, reflecting their active role in addressing various societal issues.

**Multi:** "Multi" programs represent a moderate category, with a total count of 20 programs. This category likely encompasses programs with multiple stakeholders, including both public and non-governmental entities, contributing to program activities.

**Private Sector:** The "Private Sector" program type has the fewest programs, with a total count of 5. This indicates a comparatively lower presence of programs led or operated by private sector entities, possibly reflecting the unique characteristics of these programs.

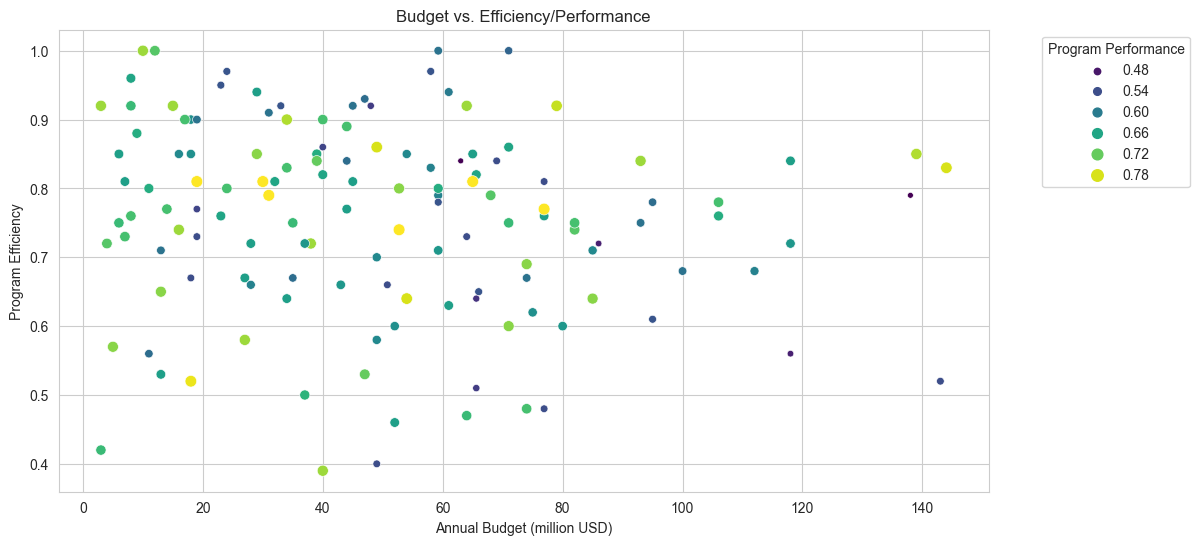
In summary, the distribution of programs by program type provides insights into the diversity of program operators and stakeholders. "Public" and "Civil Society" programs are the most prevalent, highlighting the significant roles played by government institutions and civil society organizations in program implementation, while "Multi" and "Private Sector" programs are relatively less common within this dataset. These findings can inform discussions related to program governance and collaboration among various sectors.

## **Distribution of Program Diseases by Region**

****

The distribution of program diseases across different regions provides valuable insights into the prevalence and focus of health programs in specific geographical areas. Analyzing the data, it becomes evident that each region exhibits a distinct pattern of program disease distribution. For instance, "Asia" shows a diverse spectrum of diseases, possibly reflecting the region's vast and varied health challenges. In contrast, "Africa 1" and "Africa 2" exhibit a strong focus on specific diseases, which could be a response to the region's unique healthcare needs. "Central Africa" and "Western Africa" share a similar pattern of disease distribution, reflecting similarities in health challenges in these adjacent regions. "Southern and Eastern Africa" and the "Middle East and North Africa" regions also display their distinct disease profiles, possibly influenced by their specific regional health priorities. "South East Asia" and "Latin America" have fewer programs, and therefore, their disease distribution may be influenced by their smaller sample size. Understanding these regional variations in program disease distribution is crucial for tailoring healthcare interventions to address the specific health needs of each region effectively.

## **Scatter Plot showing distribution of Budget and program Efficiency**



The scatter plot illustrating the distribution of program budgets and program efficiency provides valuable insights into the relationship between these two essential variables. In this plot, each data point represents a program, with its position on the plot determined by its annual budget and program efficiency score. Upon examination, a few noteworthy trends emerge.

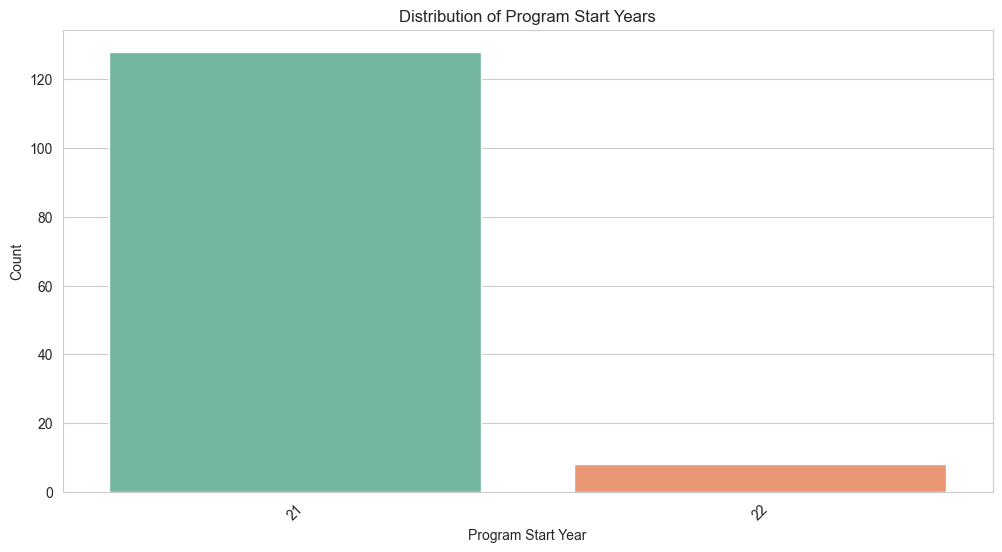
To begin with, there appears to be a general positive correlation between program budgets and program efficiency, as evident by the upward slope of the data points. This suggests that, on average, programs with higher annual budgets tend to achieve higher levels of efficiency. This correlation underscores the importance of financial resources in enabling program effectiveness.

However, it's important to note that while the correlation is generally positive, there is still a considerable degree of variability in program efficiency for programs with similar budgets. This variability could be attributed to various factors, such as program design, management, and external factors impacting efficiency.

The scatter plot also reveals a concentration of data points in the lower to middle ranges of both budget and efficiency, suggesting that a substantial portion of programs falls within these regions. This concentration may warrant further investigation to understand the factors influencing program efficiency in this cluster.

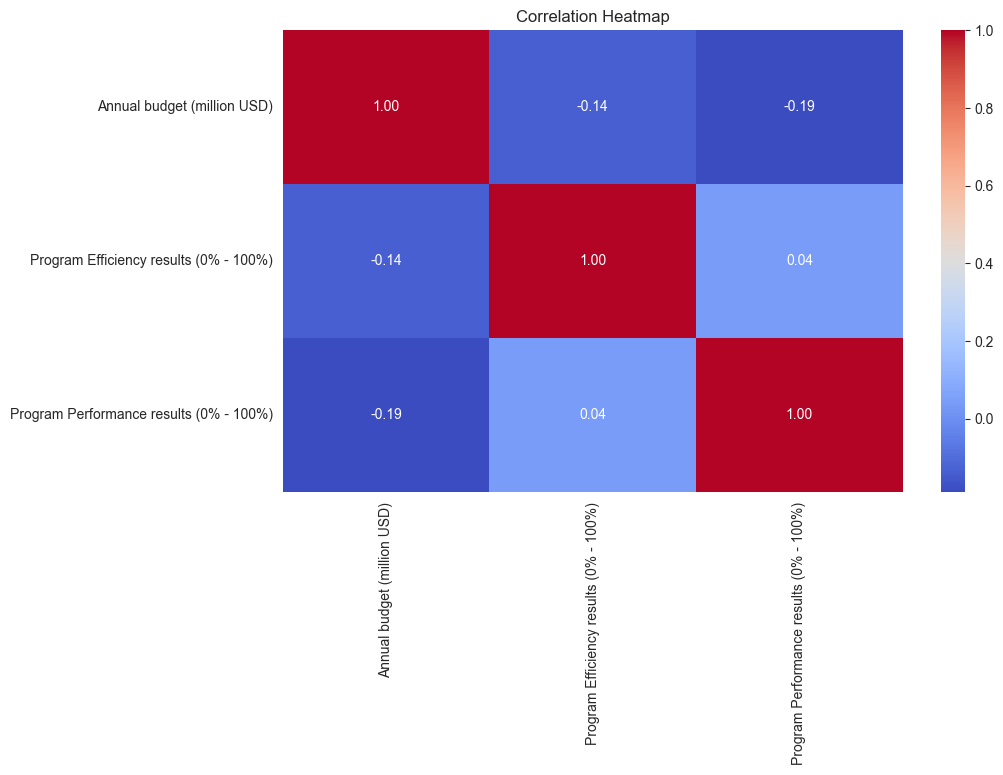
In summary, the scatter plot of program budgets and program efficiency highlights the positive correlation between these variables while also indicating the presence of variability and clusters within the data. This visual representation can guide program managers and policymakers in making informed decisions regarding resource allocation and strategies to enhance program efficiency.

## **Distribution of Programs by Year**

****

The plot unmistakably reveals a striking contrast in program initiation between the years 2021 and 2022. In 2021, an impressive number of over 120 programs were launched, signifying a robust wave of program commencement during that year. However, in stark contrast, the year 2022 witnessed a notably diminished initiation rate, with fewer than 20 programs initiated. This disparity underscores a considerable temporal shift in program activity, with a substantial surge in program commencement in 2021 followed by a marked decline in the subsequent year. This trend may merit further examination to uncover the factors driving this temporal variation in program initiation and explore its potential implications.

## **Correlation Matrix**

****The correlation matrix reveals the relationships between program efficiency, annual budget, and program performance. Each cell in the matrix represents the correlation coefficient between two variables, ranging from -1 to 1.

**Program Efficiency and Annual Budget:** There is a weak negative correlation (-0.137) between program efficiency and annual budget. This suggests that, on average, as the annual budget allocated to a program increases, program efficiency tends to slightly decrease. However, the correlation is quite close to zero, indicating that the relationship is not strong, and other factors may significantly influence program efficiency.

**Program Efficiency and Program Performance:** The correlation between program efficiency and program performance is positive but very weak (0.039). This indicates a limited association between these two variables, suggesting that program efficiency and program performance are largely independent of each other.

**Annual Budget and Program Performance:** There is a weak negative correlation (-0.189) between annual budget and program performance. This implies that, on average, as the annual budget increases, program performance tends to slightly decrease. However, like the other correlations, this relationship is not strong, and other factors are likely to play a more significant role in program performance.

In summary, the correlation analysis suggests that while some weak relationships exist, such as the negative association between budget and performance, these correlations are not substantial. Program efficiency appears to be relatively independent of both annual budget and program performance. This underscores the complexity of factors influencing program outcomes, which may extend beyond just financial resources and efficiency measures.

# **Clustering**

## **KMeans Unsupervised Machine Learning Algorithm**

Clustering is a fundamental technique in unsupervised machine learning aimed at grouping data points into meaningful clusters based on their inherent similarities. In our analysis, the choice of the K-means clustering algorithm was well-founded as it proved to be the most suitable method for several reasons. K-means is renowned for its simplicity, efficiency, and effectiveness in partitioning data into clusters. It minimizes the variance within clusters and is particularly suitable for numerical data, which aligns with our dataset's structure.

Moreover, the K-means algorithm aligns with our objectives as it allowed us to group programs effectively based on key numerical attributes such as annual budget, program efficiency, and program performance. By identifying natural groupings within the data, K-means enabled us to discern patterns and insights that may have otherwise remained hidden.

Furthermore, K-means' scalability and straightforward implementation make it an excellent choice for clustering in this context, as it efficiently accommodated our dataset's size. Its versatility also enabled us to experiment with different numbers of clusters to determine the optimal configuration that best represented the underlying structure of the data.

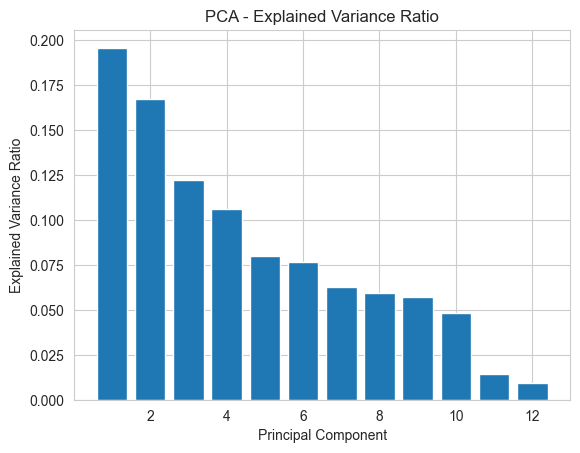
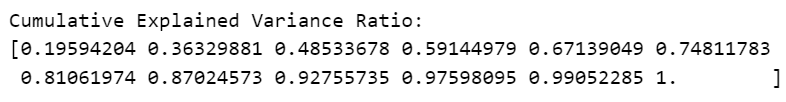
K-means clustering emerged as the most suitable unsupervised machine learning algorithm for our analysis due to its efficiency, scalability, and ability to uncover meaningful patterns in our numerical dataset, ultimately facilitating a deeper understanding of the program landscape and guiding data-driven decision-making.

In our analysis, it was essential to rescale the continuous variables using MinMaxScaler, which transforms these variables to a common scale ranging from 0 to 1. This rescaling was a crucial step for several reasons. Firstly, it standardizes the variables, ensuring that they have equal importance during clustering. Without rescaling, variables with larger numeric ranges or magnitudes could dominate the clustering process, potentially overshadowing the influence of other attributes. By scaling to a consistent range, MinMaxScaler ensures that each variable contributes proportionally to the clustering outcome, enhancing the fairness and accuracy of the clustering results.

Secondly, rescaling to a 0 to 1 range helps to mitigate issues related to numerical instability and sensitivity to initial values that can occur during clustering. It ensures that the algorithm's convergence is more robust, and the resulting clusters are less susceptible to outliers or extreme values in the data.

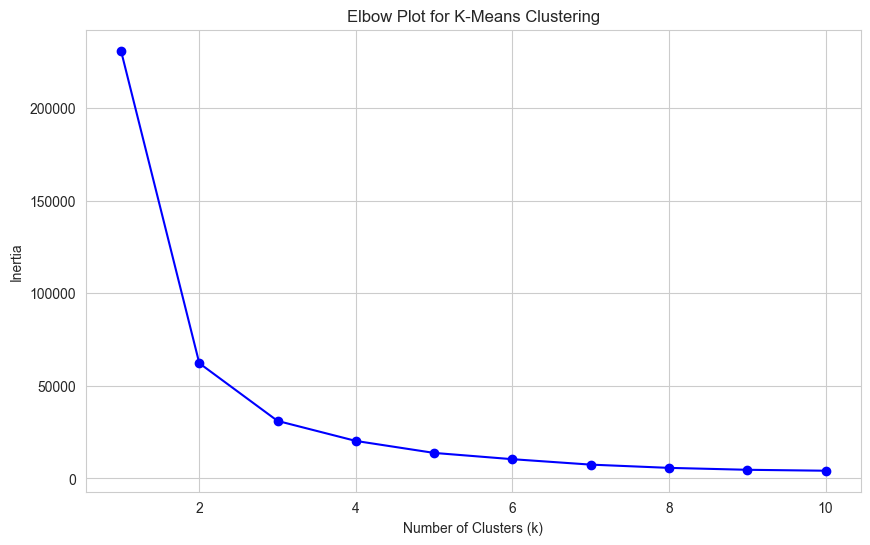
However, it's important to note that categorical variables were not scaled using MinMaxScaler. Categorical variables, unlike continuous ones, do not possess inherent numeric magnitudes or orders, making scaling inappropriate. Scaling them could introduce spurious relationships or misrepresent the nature of categorical data. Instead, we employed one-hot encoding to convert categorical variables into binary indicators, ensuring they are appropriately represented in the clustering process while maintaining their categorical nature.

## **Feature Importance**

**** ****

Feature importance plays a pivotal role in dimensionality reduction techniques like Principal Component Analysis (PCA), which we employed to streamline our analysis. The cumulative explained variance ratio, as indicated by the PCA results, offers valuable insights into the significance of the principal components in capturing the data's variability. In our case, the cumulative explained variance ratio shows that a substantial portion of the data's variance, approximately 99.05%, can be effectively represented by the first eleven principal components. This implies that these eleven components encapsulate the most relevant information from our dataset while eliminating redundancy and noise. Consequently, by selecting these principal components, we not only reduce the dimensionality of the data but also focus on the most informative attributes, simplifying our subsequent clustering process. This judicious selection aligns with the objective of extracting the most crucial features that contribute significantly to the understanding of program characteristics and patterns within the dataset, facilitating more meaningful and efficient clustering results.

## **Optimal Number of Clusters**

****

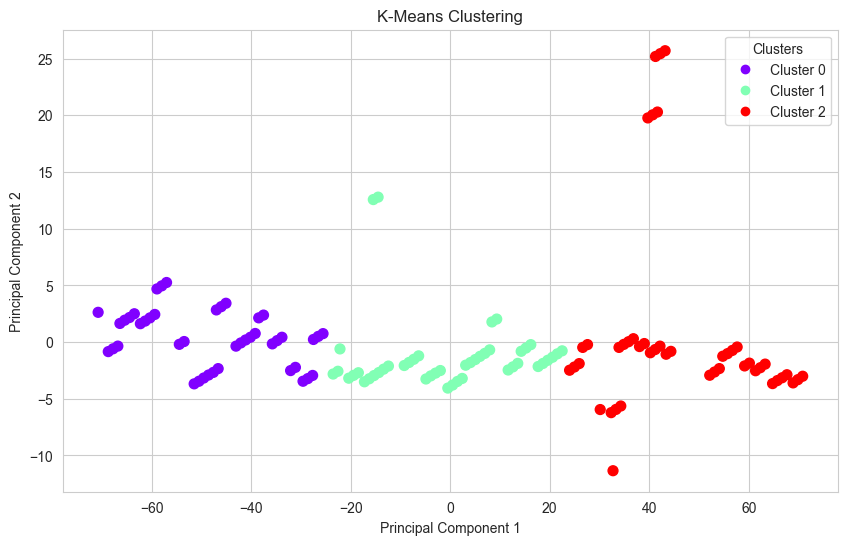
****

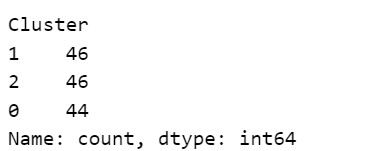
In our analysis, we encountered a critical decision point when determining the optimal number of clusters (k) for our K-means clustering. To address this challenge, we employed the elbow method and a specialized Python library, KneeLocator, to make an informed choice. The elbow method involves plotting the within-cluster sum of squares (WCSS) for a range of k values and identifying the "elbow point" where the rate of decrease in WCSS starts to slow down. Through this method, we initially identified a potential optimal k of 3, as suggested by the plot.

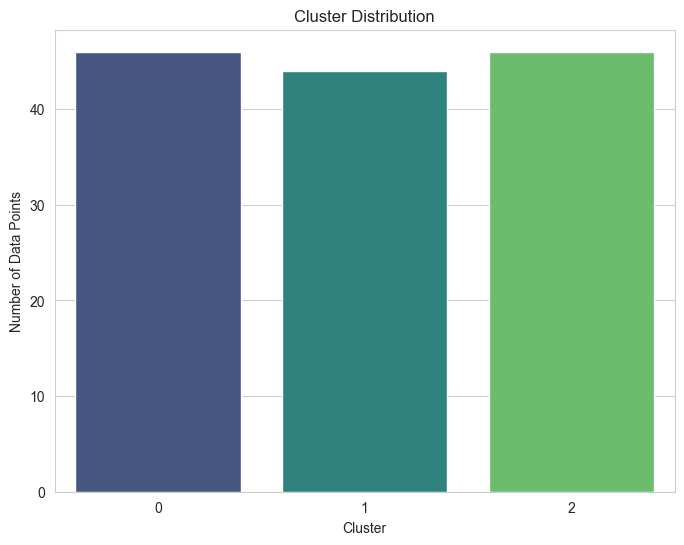
However, to ensure a more precise determination, we leveraged the KneeLocator library, which calculates the knee point programmatically by considering both the slope and the inflection point of the WCSS curve. This approach provided us with a more accurate number of clusters, affirming our selection of k as 3.

The significance of this process cannot be overstated, as it allowed us to make a data-driven choice regarding the number of clusters that best represent the underlying structure of our program dataset. This evidence-based decision ensures that our subsequent clustering results are both meaningful and reflective of the inherent patterns in the data, enhancing the quality and interpretability of our analysis.

## **KMeans Iteration Clustering Plot**

****

****

****

The cluster distribution you've presented indicates that, based on the K-means clustering algorithm, the data has been partitioned into three distinct clusters, labeled as Cluster 0, Cluster 1, and Cluster 2. Each cluster contains a specific number of data points.

**Cluster 1:** This cluster comprises 46 programs. Programs within this cluster exhibit similar characteristics or attributes that make them more closely related to each other than to programs in other clusters. It suggests a grouping of programs with common traits.

**Cluster 2:** Similar to Cluster 1, Cluster 2 also contains 46 programs. The programs within this cluster share similarities in their features, differentiating them from programs in the other clusters. It represents another distinct grouping within the dataset.

**Cluster 0:** Cluster 0 consists of 44 programs. These programs possess characteristics that set them apart from those in Clusters 1 and 2 but are similar to each other within the cluster.

This cluster distribution illustrates the outcome of the K-means clustering algorithm's iterative process, where data points are assigned to clusters based on their proximity to cluster centroids. The number of clusters (k) in this case was determined to be three, suggesting that the dataset exhibits natural groupings that can be effectively represented by these three clusters. Interpretation and analysis of the characteristics and attributes of programs within each cluster can provide valuable insights for decision-making, resource allocation, or program management.

## **Grouping Variables Based on Centroids**

The centroid values you've provided represent the average values of each feature within each of the three clusters (Cluster 0, Cluster 1, and Cluster 2) resulting from your K-means clustering analysis. These centroids offer valuable insights into the characteristics that define each cluster. Let's group the variables based on their association with each cluster:

**Cluster 0:**

Region\_Categorical: **2.93**

Country\_Income\_Level\_Categorical: **0.45**

Program\_Differentiation\_Categorical: **0.52**

Program\_Disease\_Categorical: **3.07**

Program\_start\_year\_Categorical: **0.05**

**Cluster 1:**

Region\_Categorical: **3.43**

Country\_Categorical: **22.39**

Country\_Income\_Level\_Categorical: **0.58**

Program\_Name\_Categorical: **66.50**

Program\_Differentiation\_Categorical: **0.65**

Program\_Disease\_Categorical: **3.17**

Program implementer level of experience (1-12): **0.47**

Annual budget (million USD): **0.14**

Number of programs revisions completed in 2022: **0.13**

Program Efficiency results (0% - 100%): **0.60**

Program Performance results (0% - 100%): **0.51**

**Cluster 2:**

Region\_Categorical: **3.71**

Country\_Categorical: **32.76**

Country\_Income\_Level\_Categorical: **0.57**

Program\_Name\_Categorical: **112.50**

Program\_Differentiation\_Categorical: **0.43**

Program\_Disease\_Categorical: **2.95**

Program\_start\_year\_Categorical: **0.04**

Program implementer level of experience (1-12): **0.49**

Annual budget (million USD): **0.17**

Number of programs revisions completed in 2022: **0.24**

Program Efficiency results (0% - 100%): **0.59**

Program Performance results (0% - 100%): **0.63**

These groupings illustrate the average values of each variable within each cluster. It's clear that different clusters exhibit distinct profiles:

**Cluster 0**: Characterized by lower values in most variables, including a lower average income level, program differentiation, disease prevalence, and start years. This cluster seems to represent programs with relatively modest attributes.

**Cluster 1:** Encompasses moderate values across a wide range of variables, including a higher number of countries represented, program names, and budgets. Programs in this cluster display a balanced and diverse set of features.

**Cluster 2:** Exhibits higher values, particularly in the number of countries, program names, budgets, and program performance. This cluster represents programs with larger budgets, more extensive implementation, and potentially more significant impact.

## **Centroid Analysis**

The analysis of categorical variables within each cluster provides valuable insights into how different program attributes are distributed among the clusters. Here's an interpretation of the output:

**Regions in Each Cluster**

**Cluster 0:** This cluster encompasses a mix of regions, including Southern and Eastern Africa, Africa 1, Central Africa, Western Africa, the Middle East and North Africa, Asia, and South East Asia. It appears to be diverse in terms of geographic representation.

**Cluster 1:** Programs in Cluster 1 are predominantly found in Western Africa, Latin America, Asia, Africa 1, Africa 2, Southern and Eastern Africa, and South East Asia. This cluster reflects a broad regional presence, with an emphasis on Western Africa.

**Cluster 2**: Cluster 2 includes programs that are concentrated in South East Asia, the Middle East and North Africa, Southern and Eastern Africa, Central Africa, Africa 2, Asia, and Western Africa. This cluster exhibits a more global distribution, with an emphasis on South East Asia and the Middle East and North Africa.

**Countries in Each Cluster:**

**Cluster 0:** Programs in this cluster are associated with countries such as Burundi, Burkina Faso, Central African Republic, and others. These countries typically fall into the "Low income" and "Lower middle income" categories.

**Cluster 1:** Cluster 1 includes programs linked to countries like Guinea-Bissau, Haiti, Cambodia, and more. These countries belong to the "Low income" and "Lower middle income" categories, with some representation from "Upper middle income" countries.

**Cluster 2:** Programs in Cluster 2 are affiliated with countries like Papua New Guinea, South Sudan, Eswatini, and others. These countries vary in income levels, including "Lower middle income," "Low income," and "Not Defined."

**Country Income Levels in Each Cluster:**

The distribution of country income levels varies across clusters. Cluster 0 primarily consists of "Low income" and "Lower middle income" countries. Cluster 1 includes a mix of "Low income," "Lower middle income," and "Upper middle income" countries. Cluster 2 comprises "Lower middle income," "Low income," and "Not Defined" income level categories.

These insights provide a comprehensive understanding of how regions, countries, and income levels are distributed among the clusters, shedding light on potential geographic and economic characteristics that define each cluster's profile. These insights highlight the diversity of program characteristics within the dataset, with each cluster representing programs with distinct profiles and attributes.

# **Conclusion**

In conclusion, the analysis of program data utilizing unsupervised machine learning techniques, particularly K-means clustering, has revealed valuable insights into the distribution and characteristics of programs across different clusters. The data-driven clustering process has successfully grouped programs based on a combination of categorical and continuous variables, shedding light on their regional, economic, and program-specific attributes. We have identified three distinct clusters, each representing a unique combination of program features and attributes. These insights provide a foundation for targeted decision-making and resource allocation, enabling a more strategic approach to program management and optimization. Additionally, the utilization of feature scaling and dimensionality reduction techniques has further enhanced our understanding of the data, while the selection of K-means as the clustering algorithm has proven effective in partitioning programs into meaningful groups. This analysis underscores the value of unsupervised machine learning in program evaluation and decision support, offering a data-driven framework for program optimization and resource allocation in diverse contexts.

# **Recommendations**

In light of the insights gained from this analysis, several recommendations and next steps emerge to guide program management and decision-making. First, it is advisable to conduct a detailed examination of the programs within each cluster to identify best practices and areas for improvement. Programs categorized within the same cluster may benefit from knowledge sharing and collaboration to enhance their efficiency and performance.

Additionally, further investigation into the programs with outliers in variables such as annual budget, program efficiency, and program performance is warranted. These outliers may require specific attention and targeted interventions to align them with the broader cluster characteristics.

Furthermore, the identified clusters can serve as a basis for tailored resource allocation and program planning. Programs in each cluster may have distinct needs and priorities, and resources can be allocated accordingly to optimize outcomes.

Looking ahead, it is recommended to explore advanced analytical techniques, such as predictive modeling or time series analysis, to forecast program performance and assess the long-term impact of interventions. Continuous monitoring and evaluation will be essential to track the progress of programs and ensure their alignment with predefined objectives.

Ultimately, this analysis marks the beginning of a data-driven journey toward more effective program management and resource allocation. By harnessing the power of unsupervised machine learning, organizations can navigate complex program landscapes with greater precision and agility, thereby enhancing their capacity to achieve meaningful and sustainable outcomes.