UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

Clustering Techniques in Multi-Objective Optimization: Applications in Climate-Driven Refugee Relocation

A THESIS

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

MASTER OF SCIENCE

By

Alyssa Pletcher

Norman, Oklahoma

2022

CLUSTERING TECHNIQUES IN MULTI-OBJECTIVE OPTIMIZATION: APPLICATIONS IN CLIMATE-DRIVEN REFUGEE RELOCATION

A THESIS APPROVED FOR THE

SCHOOL OF INDUSTRIAL AND SYSTEMS ENGINEERING

BY THE COMMITTEE CONSISTING OF

Dr. Kash Barker, Chair

Dr. Andres Gonzalez

Dr. Talayeh Razzaghi

Abstract

As climate change becomes increasingly concerning around the world, and with large uncertainty falling on the aspects of displaced people, a need for planning is prevalent. This complex problem—of which there is little to no preparation for—will require a comprehensive look into the different layers of the pathways to resettlement. The current process for refugee resettlement is not suitable for the prospective increase in the number of displaced people due climate related incidents, nor does it consider climate resettlement apart of the growing refugee population at the time. As this problem has proven to be laborious and extensive in the number of attributes to be considered, the goal of this study is to expand on a developing multi-objective optimization (MOO) problem by displaying how applying clustering methods can be beneficial to a resettlement plan for decision-makers. By applying k-medoids clustering (PAM) to host locations, the proposed addition aims neutralize some of the error in the arduous resettlement plan, provides the ability to adjust the granularity of focus, and takes a more practical look into an unknown, multi-faceted future.

Contents

List of Tables

List of Figures

1 Introduction

Resettlement due to environmental impacts is not a new issue. In fact, since 2008, on average, 21.5 million people have been displaced annually due to climate related incidents (United Nations High Commissioner for Refugees, 2016). Though many can be resettled amongst their own borders, the number of displaced people due to weather related natural disasters is forecasted to reach a total of 1.2 billion people by 2050 given serious action is not taken (Institute for Economics & Peace). Although climate change is not considered a valid asylum-seeking claim—thus, not qualifying for legal protection of their rights under the 1951 Refugee Convention—it is predicted to exacerbate economic, social, political, and environmental factors that cause conflict and insecurity (Werrell & Femia, 2015). That is, causing people to seek resettlement outside of their own borders. With the 25 most climate vulnerable countries, 14 are conflict riddled (International Committee of the Red Cross, 2021). Though this does not depict a correlation between the two, it suggests enduring conflict weakens the ability to adapt to climate change and conversely the impact that climate has on inciting conflict. The resolution for climate resettlement has currently been a domestic issue that studies have been conducted on but has proven itself to be an international issue with a growing number of afflicted people (World Bank Group, 2021).

1.1 Objective

As an apparent social and global issue this work aims to expand upon the planning framework from Cilali, *et al* (2021) by applying clustering methods to a version of this multi objective optimization problem. As a complex issue with little to no preparation for internationally displaced people due to climate change, this work proposes the addition of clustering to manage some of the immense uncertainty associated with this problem. By clustering the host locations goal is to alleviate some of the overcrowding issues for long-term and permanent resettlement. With clustering, the granularity of the problem can be altered—again, aiding in neutralizing some of the potential error when it comes to this arduous problem.

1.2 Literature Review

Bowerman, *et al.* (1995) introduced a multi-objective optimization approach to the school bus routing problem (SBRP). SBRP, prior to this study, used density-based clustering techniques to group users to potential bus stops, and then employed a traveling salesman problem algorithm to route the drivers. The proposed formulation expanded on a traditional routing problem by considering equity, efficiency, and cost, as well as walking distance and driving distance. This study additionally proposes the use of a user defined weighting system, such that the publicly funded transportation routes can adjust the weight of importance on the clustering criterion. As the decision-maker had more influence on the results, Bowerman, *et al.* (1995) show the benefits in the use of clustering in the MOO model as the decision makers could modify the clusters while maintaining the integrity of the optimization problem.

In a 2019 study conducted by Wang, *et al.* (2012), clustering techniques were employed in identifying the realizations of a retrospective optimization problem to account for geological uncertainty. The proposed study focused on identifying the optimal location for drilling wells. They introduced a retrospective optimization (RO) model such the framework would consider both stochastic and deterministic core optimizers. As the algorithm generated a multitude of realizations, Wang, et al. (2012) approached their selection with Random selection and with a k-means clustering algorithm for selection. Utilization of clustering techniques resulted in a higher cumulative oil production with few simulation calls and fewer iterations of the RO model. This method, though not applied to this optimization problem in the same manner, shows another benefit of the implementation of clustering in a multi objective optimization problem.

Mousa *et. al* (2017) proposes the use of clustering to reduce the population for resource allocation with the goal to minimize cost and maximize efficiency. The use of clustering allowed different genetic algorithm operators to be applied to sub-populations (clusters) such that the clusters could be dynamic and diverse. When assessing the problem with a singular population, only one genetic algorithm operator could be used. This study concluded that the solution sets obtained when using clustering data is dominant to those obtained by using the whole population.

2 Methodology

As an extension to the work conducted by Cilali, *et al.*, in addition to the clustering algorithm proposed for host locations, a portion of this work includes a variation of the facility location and allocation model (2021).

2.1 Clustering

When dealing with a global issue like climate change, the “red tape” that needs to be considered for relocating people across international borders can heavily affect the planning process. By clustering locations (whether clustering by city, country, region, etc.), the process of relocation can become more stratified allowing decision makers to focus on the necessary problem (e.g., by clustering countries, and appropriately relocating people to the clusters, the problem can then additional iterations can be observed within the cluster.

As this problem deals with preexisting work on mass relocation of communities affected by climate change, the goal is to alleviate work for decision makers by giving them multiple host locations rather than one as well as in the study by allowing for a more heuristic approach.

2.1.1 Attributes

In the selection of attributes to represent the countries, it was prioritized to include those factors that most influence a refugee's integration into a new society. Although it is not an exact science, there should be significance given to the comfortability of those needing resettlement, as proposed in Cilali, *et al* (2021).

A look into the manner in which refugees are being assigned to host locations was conducted. Conflict refugees are forced to flee their home country due to fear of persecution or conflict. Though the qualities that a person fears persecution for could provide an interesting cluster of potential host countries, this still does not consider diversity. Grouping countries based on these attributes would lead to a large influx of an already dominant social/demographic group. According to Chris Boian, a spokesman for the UNHCR, the process of resettlement, for those who agree to enter and are deemed in need, begins with background check and interviews collecting biographical information and biometric data; this is then used to place refugees with a host country by taking into consideration the host countries quota, the refugees' relationships to residents in the host country, and their cultural affinities (Epatko, 2017). It will be assumed that climate refugees are not necessarily seeking a host location for a change in the cultural society, rather, they are seeking host locations much like their own, just out of the threat of climate—this does not mean that all resettled people would not prefer to be resettled in a country with more opportunity, as societies are not homogenous culturally. This leave us with the question: how are countries defined and what makes them similar? It is proposed that the attributes that are used to cluster countries include those that are fully representative of the location (i.e., not just social, but physical, environmental, occupational, and so on). This can be generalized as data that is represented by the standard of living (quantifiable) as an alternative to quality of life (more subjective features that are not necessarily measurable).

2.1.2 Data

The data to conduct the clustering was collected from a country comparison option via the CIA Factbook (*The World Factbook 2021*.). This almanac style reference resource contains information and statistics over different regions, countries, and territories around the world. In adhering to a standard of living focus, the attributes that were chosen from the empirical data site are representative of the flowing topics:

- Geography
- People and society
- Environment
- Economy
- Energy
- Communication
- Transportation
- Military and Security

The attributes can be seen listed in Table 4 within Appendix A.

2.1.3 Exploratory Data Analysis

To decide on the appropriate clustering algorithm, an exploratory data analysis was conducted to view the essential characteristics of the data. This data frame consisted of 257 locations along 54 attributes. This included 2653 missing values. In viewing a description of the attributes, it was observed that the was a relatively high variance among the observations. As well as a relationship between the variance and the type of statistic (i.e., there was high variance in variables falling under the economy category).

Based on this full analysis (see Appendix B) centroid-based clustering method was chosen. This is due to the potentially high computational effort, and, more specifically, k-medoids or Partitioning Around Mediods (PAM) algorithm was selected dur to the substantial number of outliers that were present for each variable (Figure 1).

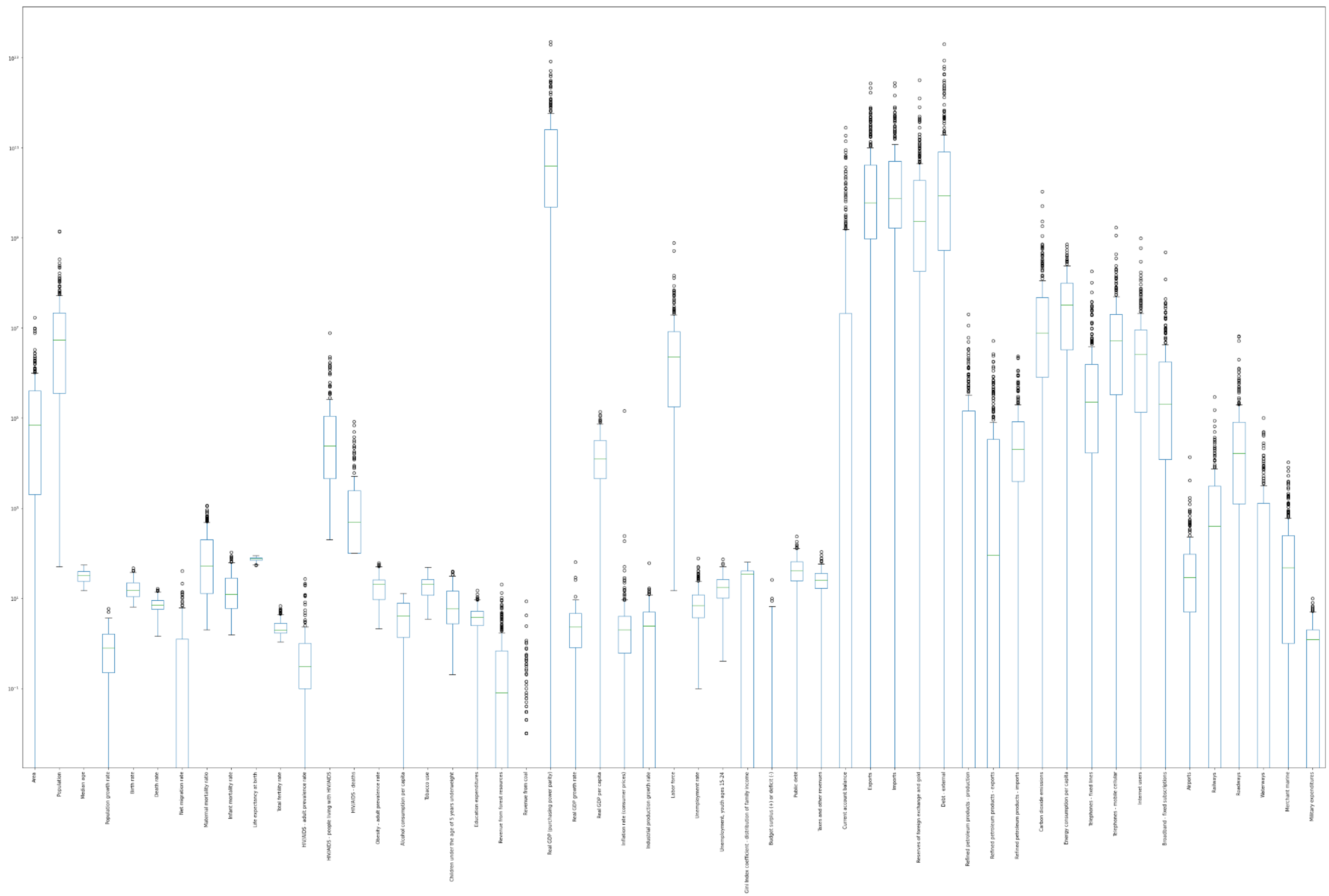


Figure 1. Box plots for clustering attributes

2.1.4 Data Preparation

A diagnostic of the missing data mechanism must be done as PAM cannot have missing values (see Appendix B). Based on domain knowledge and an analyzation of the data descriptions, most data points were assumed to be missing not at random (i.e., the missingness was due to other circumstances). if data is missing it is because it is not available either because it was not recorded, released, or possible to obtain in a particular location to record it was assumed to be 0. Additionally, some variables are identified to be missing as they are assumed to be meaningful zeros. For example, for the value of Airports, the description of the attributes obtained (seen in Appendix A), airports are described as those airports of airfields recognizable from the air. It can be inferred if the data point was missing that is because there were none to be seen from air. Therefore, the data is non-ignorable and must be modeled or removed. The remaining values that were missing at random, a k-nearest neighbors' imputation was used.

2.2 Application of Clustering: Implementation into Multi-Objective Model

This section includes a variation of the model proposed in Cilali *et. al* (2017) with the addition of clusters as host locations. This optimization model aims to minimize the costs and number of individuals that failed to be resettled.

2.2.1 Assumptions

The following assumptions were made for the formulation of this model:

- All countries with a natural inhabitant population agree with other members of their cluster to participate in accepting refugees. This includes no conflicting interests of the members in the cluster for the individuals resettled within the cluster.

- Those in need of resettlement will be considered individuals, rather than groups or family units. Therefore, the capacity, flow from origin to host, and demand are measured with respect to the number of individuals.
- Within clusters, all parties will assume the same responsibilities in terms of opening to accept refugees, opening costs, and take on the decision for resettlement within the cluster.
- The likelihood of occurrence for each demand scenario is deterministic.
- The costs for opening a host cluster, cost of expansion, and cost of deviation are known for each cluster and were agreed upon by members of the cluster and are available and known at the time of running this model. Additionally, these values are the summation of each cluster members respective values (e.g., the opening cost for one cluster is the summation of the opening costs for each member within the cluster). This is a relevant assumption to be made for the comparison of the model with and without clustering.
- For this model, cost is assumed in relation to the number of individuals (e.g., opening cost is the number of individuals that one cluster can take on).
- The upper bound for the number of individuals that one cluster can take on in one time period is unchanged for each subsequent time period and resets at each new time period.
- The origin locations are assumed to be countries vulnerable to climate change but is not inclusive of all locations vulnerable to climate change.
- The demand scenarios were assumed to be deterministic.

2.2.2 Sets

The model aims to relocate individuals from origin locations, to host clusters, under 4 different time periods, under 3 deterministic scenarios.

Table 1. Sets and indices.

I	Set of origin locations, indexed by i
J	Set of host clusters, indexed by j
T	Set of time periods, indexed by t
Ω	Set of scenarios, indexed by ω

The origin locations for this model includes 29 different locations considered to be vulnerable for climate change. The host clusters were created using country data which resulted in 4 different clusters. Time periods for this model represent 5-year increments and the set of scenarios. The set of scenarios include low, medium, and high. That is low being not as many inhabitants as predicted are in need of resettlement, medium being the forecasted number of individuals in need of resettlement is as is, and high is more than predicted are in need of resettlement.

2.2.3 Parameters

The parameters for this model are primarily assumed for the model, as obtaining the values would include involvement from relevant decision makers. Each cluster has its own opening cost and initial capacity (for this model this is equivalent to the number of individuals the cluster can take in initially). The opening cost is only applied once as a cluster begins accepting refugees. There is also a cost associated with exceeding this value and an upper bound on the capacity change a host location can take on during one period. Each origin location has a remand associated with it for different time periods under different scenarios. These scenarios have an assumed deterministic probability associated with them, that is, there is an equal likelihood that any scenario can take place. In this case, it is assumed there are low, medium, and high demand scenarios are represented by 50%, 100%, and 150%, respectively, of the vulnerable population that need resettlement.

Due to the urgency of the relocation, there is also a cost associated with the inability to reach a relocation demand.

Table 2. Parameters.

UB^{φ}	Highest amount of capacity change allowed for any host location within one period
c_j^a	Cost of opening host cluster $j \in J$ for the first time
c_j^e	Cost of expansion host cluster $j \in J$
cd	Unit cost of deviating from the relocation goal by one individual
Φ_{j0}	Initial resettlement capacity of the host location $j \in J$ at period zero
p^{ω}	Probability of scenario $\omega \in \Omega$
d_{it}^{ω}	Relocation demand forecasted for origin location $i \in I$ at period $t \in T$ under scenario $\omega \in \Omega$

2.2.4 Decision Variables

The values we are looking to obtain in this model includes the number of individuals relocating from an origin to a cluster at different time periods under each scenario. This value ideally lays out the plan for relocation of individuals. Other important decision variables include the binary value opening of a host location at each time period, as the decision makers will know the period to which they need to prepare for accepting refugees.

Additionally, the capacity expansion value is imperative for preparation. The total resettlement capacity is a value that will be decided by the model to adjust how much the host cluster will need to expand for the upcoming period—with the upper limit noted in parameters. Additionally, the resettlement demand of the origin locations will likely fluctuate based on how many individuals relocated in previous periods and the different scenarios.

Table 3. Decision Variables.

y_{jt}	Binary variable equal to 1 if host location $j \in J$ is opened at period $t \in T$, 0 otherwise
$\Delta\varphi_{jt}$	Capacity expansion for the upcoming period of the host location $j \in J$ at period $t \in T$
φ_{jt}	Total resettlement capacity of the host location $j \in J$ at the beginning of period $t \in T$
x_{ijt}^{ω}	Number of individuals assigned / relocation flow from origin location $i \in I$ to host location $j \in J$ at period $t \in T$ under each scenario $\omega \in \Omega$
γ_{it}^{ω}	Resettlement demand of the origin location $i \in I$ at the beginning of period $t \in T$ under each scenario $\omega \in \Omega$

2.2.5 Objective Functions

The objective functions are formulated in Eqs. (1 – 4). These objective functions represent both cost and demand. Eq. (1) represents the opening costs, Eq. (2) represents the expansion costs, and Eq. (3) represents the failure to meet demand. All of which need to be minimized to incentivize participation in the resettlement process. Eq. (4) aims to minimize the deviation from the relocation goals under all possible scenarios. This is such that all individuals that need to be resettled can be resettled if resources are available.

$$\min \qquad \sum_{t\in T}\sum_{j\in J} c_j^a \times y_{jt} \tag{1}$$

$$\min \qquad \sum_{t\in T}\sum_{j\in J} c_j^e \times \Delta\varphi_{jt} \qquad - \tag{2}$$

$$\min \qquad \sum_{w\in\Omega} p^\omega \left(\sum_{i\in I}\sum_{j\in J}\sum_{t\in T} cd \quad \times x_{ijt}^\omega\right) \tag{3}$$

$$\min \qquad \sum_{w\in\Omega} p^\omega \left(\sum_{i\in I}(\gamma_{i|T|}^\omega - \sum_{j\in J} x_{ij(|T|)}^\omega)\right) \tag{4}$$

2.2.6 Constraints

Eq. (5) serves as a limit on the number of times a host location can be opened. They can only open once over all time periods.

$$\text{s.t.} \qquad \sum_{t\in T} y_{jt} \leq 1 \qquad \forall\, j \in J \tag{5}$$

Eq. (6) initializes the capacity and Eq. (7) ensure the capacity of a host cluster during a time period is adjusted to the previous period's capacity and the capacity expansion while Eq. (8) ensures this capacity expansion does not exceed the upper limit on capacity expansion. Eq. (9) limits the flow from exceeding the set capacity during a period because its expansion was calculated in the previous period.

$$\varphi_{j1} = \Phi_{j0} \qquad \forall\, j \in J \tag{6}$$

$$\varphi_{jt} = \varphi_{j(t-1)} + \Delta\varphi_{j(t-1)} \qquad \forall\, j \in J, t \in T\backslash\{1\} \tag{7}$$

$$\Delta\varphi_{j(t-1)} \leq \sum_{m=1}^{t-1} y_{jm} \times UB^{\varphi} \qquad \forall\, j \in J, t \in T\backslash\{1\} \qquad (8)$$

$$\sum_{m=1}^{t} \sum_{i \in I} x_{ijt}^{\omega} \leq \sum_{m=1}^{t} y_{jm} \times \varphi_{jt} \qquad \forall\, j \in J, t \in T, \omega \in \Omega \qquad (9)$$

The flow also cannot be greater than the demand or the number of individuals in need of resettlement in a period (Eq. 10) This demand is initialized in the first time with the forecasted demand (Eq 11). The resettlement demand for time periods is set to the demand of the previous periods are assumed to be the previous periods demand with the addition of the forecasted demand, minus those who were previously resettled (Eq. 12).

$$\sum_{j \in J} x_{ijt}^{\omega} \leq \gamma_{it}^{\omega} \qquad \forall\, i \in I, t \in T, \omega \in \Omega \qquad (10)$$

$$\gamma_{i1}^{\omega} = d_{i1}^{\omega} \qquad \forall\, i \in I, \omega \in \Omega \qquad (11)$$

$$\gamma_{it}^{\omega} = \gamma_{i(t-1)}^{\omega} - \sum_{j \in J} x_{ij(t-1)}^{\omega} + d_{it}^{\omega} \qquad \forall\, i \in I, t \in T\backslash\{1\}, \omega \in \Omega \qquad (12)$$

The final constraints, Eq. (13) and Eq. (14), are integer constraints and binary constraints, respectively.

$$x_{ijt}^{\omega}, \Delta\varphi_{jt}, \varphi_{jt}, \gamma_{it}^{\omega} \in \mathbb{Z}_{\geq 0} \qquad (13)$$

$$y_{jt} \in \mathbb{Z}_2 \qquad (14)$$

3 Results

3.1 Clustering

For PAM, the number of clusters must be input into the algorithm. Based on the within clusters sum of squares values, the number of clusters was selected as 4 (Fig. 2).

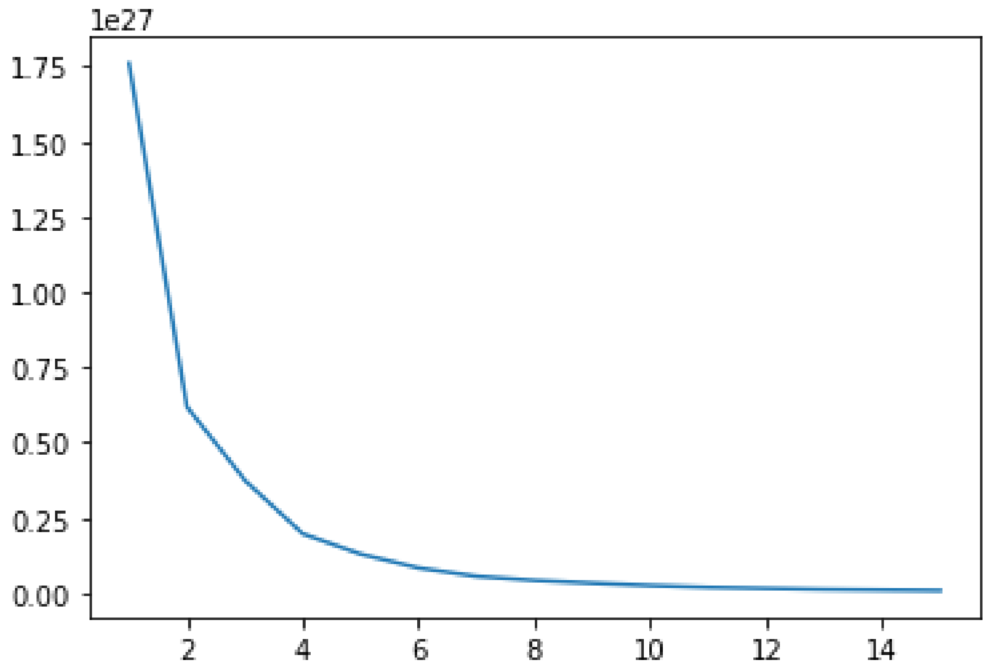


Figure 2. Within clusters sum of squares values for selecting number of clusters.

The tables below depict the number of locations in each cluster (Table 4) and the number of locations that fall in each cluster grouped by region (Table 5).

Table 4. Number of locations in each cluster

Cluster	Locations
1	78
2	74
0	49
3	36

Table 5. Number of locations in each cluster by region.

Region	Cluster	Locations
Africa	0	43
	1	1
	2	9
	3	2
Australia and Oceania	0	1
	1	3
	2	18
	3	4

Central America and the Caribbean	1	15
	2	11
	3	6
Central Asia	1	1
	2	3
	3	2
East and Southeast Asia	0	1
	1	7
	2	3
	3	10
Europe	1	45
	2	4
	3	2
Middle East	0	2
	1	1
	2	15
	3	1
North America	1	4
	2	2
South America	1	1
	2	7
	3	5
South Asia	0	2
	2	2
	3	4

Each cluster has an ample number of participants in them aiding in the goal to have option within the cluster for relocation. Though it wasn't necessary to make them density based, the clusters needed to ideally have more than a singular option as that would defeat the need for clustering. This was achieved using PAM.   We can see that the Africa predominantly makes up one cluster, while the "Western" (majority of North America, Europe, and Central America) region locations dominate another cluster. This makes some intuitive sense as the region seemed to correlate with some of the standard of living variables.

The map below provides a better visual aid for the clusters (Fig. 3). We can see that the clusters resulted have locations in almost all regions of the world, such that the clusters wasn't predominately influenced by region.

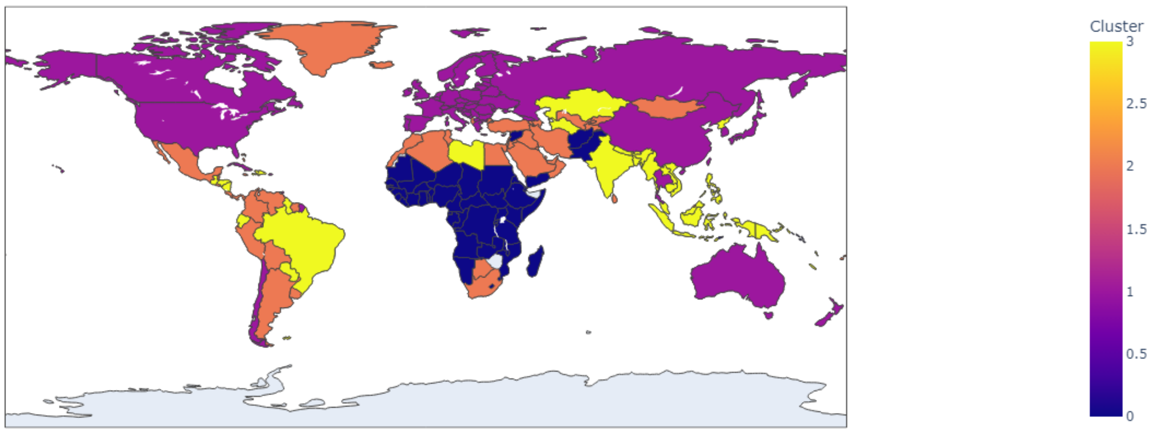


Figure 3. Map of clusters.

Some other interesting characteristics of the clusters include that cluster 0 has the lowest median CO^2 emissions while cluster 1 has the highest GDP. Cluster 2 has the highest median educational expenditures and the cluster 3 median falls somewhere in the middle for these features.

3.2 Refugee Relocation Experimental Results

The weighted sum method was applied to the multiple objectives. Because all decision variables we in terms of individuals, the values do not need to be scaled.

To compare the model with clustering to a model without clustering, the values parameters were adjusted to split between x number of locations. For example, the value for a clusters cost of opening, cost of expansion, and capacities will be divided up to represent the individual members of the clusters.

When the model was applied to the cluster, based on the parameters that were assigned (see Appendix B) the model takes on objective values for both the opening cost and the and the cost of deviation. The model applied to non-clustered locations additionally took on values for the same objective values. Below are the results of each individual objective value given the PAM was applied and when the locations were not clustered (Table 6).

Table 6. Objective Values for both alternatives

	Clusters	**No Clusters**
Eq. (1)	349840000	282336000
Eq. (3)	3953160000	14683296000

Eq. (1) represents the minimized cost of opening for which modeling without clusters produced a lower value. However, with the lack of clusters, a much higher cost of deviating from the relocation goal was produced (Eq. 3).

4 Conclusion

The intentions of the study were to evaluate the benefits of clustering techniques applied to MOO problems. The results of this study were unfortunately determined to be inconclusive as the production of a viable optimal solution was not produced. Thus, resulting in a failure to compare the use of multi-objective models with and without clustering algorithms applied. Upon further investigation, it is likely that failure lied within the interpretation of the model proposed by Cilali *et. al* and not with the introduction of clusters to the extensive look into the process of refugee resettlement due to climate change. However, as the model comes to fruition, this study serves as a place of reevaluation for the creation of a more feasible model to address the climate change epidemic.

4.1 Future Work

To further show the benefit of clustering in MOO, a stochastic approach to the model would provide better insight on the uncertainty that climate change to alleviate the unpredictable demand and capacity for relocation due to climate change climate change. Additionally, as the model becomes fully functional, the implementation of representative data be used to see the palpability of the model with clusters.

5 References

Bowerman, R., Hall, B., & Calamai, P. (1995). A multi-objective optimization approach to urban school bus routing: Formulation and solution method. Transportation Research Part A: Policy and Practice, 29(2), 107-123.

Cilali, B., Barker, K., & González, A. D. (2021). A location optimization approach to refugee resettlement decision-making. Sustainable Cities and Society, 74, 103153. https://doi.org/10.1016/j.scs.2021.103153

Climate change and conflict. International Committee of the Red Cross. (2022, March 22). From https://www.icrc.org/en/what-we-do/climate-change-conflict

Epatko, L. (2017, March 1). You asked: How are refugees referred to live in the U.S.? PBS. From https://www.pbs.org/newshour/world/asked-refugees-referred-live-u-s

Institute for Economics & Peace (IEP). (n.d.). IEP: Over one billion people at threat of being displaced by 2050 due to environmental change, conflict and civil unrest. From, https://www.prnewswire.com/news-releases/iep-over-one-billion-people-at-threat-of-being-displaced-by-2050-due-to-environmental-change-conflict-and-civil-unrest-301125350.html

International Committee of the Red Cross. (2021, October 19). Seven things you need to know about climate change and conflict. International Committee of the Red Cross. From https://www.icrc.org/en/document/climate-change-and-conflict

Likas, A., Vlassis, N., & Verbeek, J. J. (2002, May 14). The global K-means clustering algorithm. Pattern Recognition. From https://www.sciencedirect.com/science/article/pii/S0031320302000602

Mousa, A. A., El-Shorbagy, M. A., & Farag, M. A. (2017). K-means-clustering based evolutionary algorithm for multi-objective resource allocation problems. Appl. Math. Inf. Sci, 11(6), 1681-1692.

The World Factbook 2021. Washington, DC: Central Intelligence Agency, 2021. https://www.cia.gov/the-world-factbook/

United Nations High Commissioner for Refugees. (2016, November 6). Frequently asked questions on climate change and disaster displacement. From https://www.unhcr.org/uk/news/latest/2016/11/581f52dc4/frequently-asked-questions-climate-change-disaster-displacement.html

Wang, H., Ciaurri, D. E., Durlofsky, L. J., & Cominelli, A. (2012). Optimal well placement under uncertainty using a retrospective optimization framework. Spe Journal, 17(01), 112-121.

Waters, M. C., & Jiménez, T. R. (2005). Assessing immigrant assimilation: New empirical and theoretical challenges. Annual Review of Sociology, 31(1), 105–125. https://doi.org/10.1146/annurev.soc.29.010202.100026

Werrell, C. E., & Femia, F. (2015). Climate Change as Threat Multiplier: Understanding the Broader Nature of the Risk.

World Bank Group. (2021, September 13). Groundswell: Acting on Internal Climate Migration. World Bank. From https://www.worldbank.org/en/news/feature/2021/09/13/millions-on-the-move-in-their-own-countries-the-human-face-of-climate-change

Appendix A: Attribute Log for Clustering

Table 7. Attribute Descriptions

Attribute	Missingness Mechanism	Description
Airports	MNAR	"This entry gives the total number of airports or airfields recognizable from the air. The runway(s) may be paved (concrete or asphalt surfaces) or unpaved (grass, earth, sand, or gravel surfaces) and may include closed or abandoned installations. Airports or airfields that are no longer recognizable (overgrown, no facilities, etc.) are not included. Note that not all airports have accommodations for refueling, maintenance, or air traffic control."
Alcohol consumption per capital	MAR	This entry provides information on alcohol consumption per capita (APC), which is the recorded amount of alcohol consumed per capita by persons aged 15 years and over in a calendar year, measured in liters of pure alcohol. APC is broken down further into beer, wine, spirits, and other subfields. Beer includes malt beers, wine includes wine made from grapes, spirits include all distilled beverages, and other includes one or several other alcoholic beverages, such as fermented beverages made from sorghum, maize, millet, rice, or cider, fruit wine, and fortified wine. APC only takes into account the consumption that is recorded from production, import, export, and sales data, primarily derived from taxation.
Area	MNAR	"This entry gives the total number of airports or airfields recognizable from the air. The runway(s) may be paved (concrete or asphalt surfaces) or unpaved (grass, earth, sand, or gravel surfaces) and may include closed or abandoned installations. Airports or airfields that are no longer recognizable (overgrown, no facilities, etc.) are not included. Note that not all airports have accommodations for refueling, maintenance, or air traffic control."
Birth rate	MAR	"This entry gives the average annual number of births during a year per 1,000 persons in the population at midyear; also known as crude birth rate. The birth rate is usually the dominant factor in determining the rate of population growth. It depends on both the level of fertility and the age structure of the population."

Broadband - fixed subscriptions	MNAR	"This entry gives the total number of fixed-broadband subscriptions, as well as the number of subscriptions per 100 inhabitants. Fixed broadband is a physical wired connection to the Internet (e.g., coaxial cable, optical fiber) at speeds equal to or greater than 256 kilobits/second (256 kbit/s)."
Budget surplus (+) or deficit (-)	MNAR	"This entry records the difference between national government revenues and expenditures, expressed as a percent of GDP. A positive (+) number indicates that revenues exceeded expenditures (a budget surplus), while a negative (-) number indicates the reverse (a budget deficit). Normalizing the data, by dividing the budget balance by GDP, enables easy comparisons across countries and indicates whether a national government saves or borrows money. Countries with high budget deficits (relative to their GDPs) generally have more difficulty raising funds to finance expenditures, than those with lower deficits."
Carbon dioxide emissions	MNAR	"This field refers to a country's amount of carbon dioxide released by burning coal, petroleum, and natural gas. Data are reported in metric tonnes of CO2."
Children under the age of 5 years underweight	MAR	"This entry gives the percent of children under five considered to be underweight. Underweight means weight-for-age is less than minus two standard deviations from the median of the World Health Organization Child Growth Standards among children under 5 years of age. This statistic is an indicator of the nutritional status of a community. Children who suffer from growth retardation as a result of poor diets and/or recurrent infections tend to have a greater risk of suffering illness and death."
Current account balance	MNAR	"This entry records a country's net trade in goods and services, plus net earnings from rents, interest, profits, and dividends, and net transfer payments (such as pension funds and worker remittances) to and from the rest of the world during the period specified. These figures are calculated on an exchange rate basis, i.e., not in purchasing power parity (PPP) terms."

Death rate	MAR	"This entry gives the average annual number of deaths during a year per 1,000 persons at midyear; also known as crude death rate. The death rate, while only a rough indicator of the mortality situation in a country, accurately indicates the current mortality impact on population growth. This indicator is significantly affected by age distribution, and most countries will eventually show a rise in the overall death rate, in spite of continued decline in mortality at all ages, as declining fertility and increased lifespans result in an aging population."
Debt - external	MNAR	"This entry gives the total public and private debt owed to nonresidents repayable in internationally accepted currencies, goods, or services. These figures are calculated on an exchange rate basis, i.e., not in purchasing power parity (PPP) terms."
Education expenditures	MNAR	"This entry provides the public expenditure on education as a percent of GDP."
Energy consumption per capita	MNAR	"This entry refers to a country's total energy consumption per capita, including the consumption of petroleum, dry natural gas, coal, net nuclear, hydroelectric, and non-hydroelectric renewable electricity. Data are reported in British thermal Units per person (Btu/person)."
Exports	MNAR	"This entry provides the total US dollar amount of merchandise exports on an f.o.b. (free on board) basis. These figures are calculated on an exchange rate basis, i.e., not in purchasing power parity (PPP) terms."
Gini Index coefficient - distribution of family income	MNAR	"This entry measures the degree of inequality in the distribution of family income in a country. The index is calculated from the Lorenz curve, in which cumulative family income is plotted against the number of families arranged from the poorest to the richest. The index is the ratio of (a) the area between a country's Lorenz curve and the 45 degree helping line to (b) the entire triangular area under the 45 degree line. The more nearly equal a country's income distribution, the closer its Lorenz curve to the 45 degree line and the lower its Gini index, e.g., a Scandinavian country with an index of 25. The more unequal a country's income distribution, the farther its Lorenz curve from the 45 degree line and the higher its Gini index, e.g., a Sub-Saharan country with an index of 50. If income were distributed with perfect equality, the Lorenz curve would coincide with the 45 degree line and the index

		would be zero; if income were distributed with perfect inequality, the Lorenz curve would coincide with the horizontal axis and the right vertical axis and the index would be 100."
HIV/AIDS - adult prevalence rate	MAR	"This entry gives an estimate of the percentage of adults (aged 15-49) living with HIV/AIDS. The adult prevalence rate is calculated by dividing the estimated number of adults living with HIV/AIDS at yearend by the total adult population at yearend."
HIV/AIDS - deaths	MAR	"This entry gives an estimate of the number of adults and children who died of AIDS during a given calendar year."
HIV/AIDS - people living with HIV/AIDS	MAR	"This entry gives an estimate of all people (adults and children) alive at yearend with HIV infection, whether or not they have developed symptoms of AIDS."
Imports	MNAR	"This entry provides the total US dollar amount of merchandise imports on a c.i.f. (cost, insurance, and freight) or f.o.b. (free on board) basis. These figures are calculated on an exchange rate basis, i.e., not in purchasing power parity (PPP) terms."
Industrial production growth rate	MNAR	"This entry gives the annual percentage increase in industrial production (includes manufacturing, mining, and construction)."
Infant mortality rate	MAR	"This entry gives the number of deaths of infants under one year old in a given year per 1,000 live births in the same year. This rate is often used as an indicator of the level of health in a country."
Inflation rate (consumer prices)	MNAR	"This entry provides the annual inflation rate, as calculated by the percent change in current consumer prices from the previous year's consumer prices."

Internet users	MNAR	"This entry gives the total number of individuals within a country who can access the Internet at home, via any device type (computer or mobile) and connection. The percent of population with Internet access (i.e., the penetration rate) helps gauge how widespread Internet use is within a country. Statistics vary from country to country and may include users who access the Internet at least several times a week to those who access it only once within a period of several months."
Labor force	MAR	"This entry contains the total labor force figure."
Life expectancy at birth	MAR	"This entry contains the average number of years to be lived by a group of people born in the same year, if mortality at each age remains constant in the future. Life expectancy at birth is also a measure of overall quality of life in a country and summarizes the mortality at all ages. It can also be thought of as indicating the potential return on investment in human capital and is necessary for the calculation of various actuarial measures."
Maternal mortality ratio	MAR	"The maternal mortality ratio (MMR) is the annual number of female deaths per 100,000 live births from any cause related to or aggravated by pregnancy or its management (excluding accidental or incidental causes). The MMR includes deaths during pregnancy, childbirth, or within 42 days of termination of pregnancy, irrespective of the duration and site of the pregnancy, for a specified year."
Median age	MAR	"This entry is the age that divides a population into two numerically equal groups; that is, half the people are younger than this age and half are older. It is a single index that summarizes the age distribution of a population. Currently, the median age ranges from a low of about 15 in Niger and Uganda to 40 or more in several European countries and Japan. See the entry for "Age structure" for the importance of a young versus an older age structure and, by implication, a low versus a higher median age."

Merchant marine	MNAR	"This entry provides the total and the number of each type of privately or publicly owned commercial ship for each country; military ships are not included; the five ships by type include: bulk carrier - for cargo such as coal, grain, cement, ores, and gravel; container ship - for loads in truck-size containers, a transportation system called containerization; general cargo - also referred to as break-bulk containers - for a wide variety of packaged merchandise, such as textiles, furniture and machinery; oil tanker - for crude oil and petroleum products; other - includes chemical carriers, dredgers, liquefied natural gas (LNG) carriers, refrigerated cargo ships called reefers, tugboats, passenger vessels (cruise and ferry), and offshore supply ships "
Military expenditures	MNAR	"This entry gives estimates on spending on defense programs for the most recent year available as a percent of gross domestic product (GDP). For countries with no military forces, this figure can include expenditures on public security and police."
Net migration rate	MNAR	"This entry includes the figure for the difference between the number of persons entering and leaving a country during the year per 1,000 persons (based on midyear population). An excess of persons entering the country is referred to as net immigration (e.g., 3.56 migrants/1,000 population); an excess of persons leaving the country as net emigration (e.g., -9.26 migrants/1,000 population). The net migration rate indicates the contribution of migration to the overall level of population change. The net migration rate does not distinguish between economic migrants, refugees, and other types of migrants nor does it distinguish between lawful migrants and undocumented migrants."
Obesity - adult prevalence rate	MAR	"This entry gives the percent of a country's population considered to be obese. Obesity is defined as an adult having a Body Mass Index (BMI) greater to or equal to 30.0. BMI is calculated by taking a person's weight in kg and dividing it by the person's squared height in meters."

Population	MNAR	"This entry gives an estimate from the US Bureau of the Census based on statistics from population censuses, vital statistics registration systems, or sample surveys pertaining to the recent past and on assumptions about future trends. The total population presents one overall measure of the potential impact of the country on the world and within its region. Note: Starting with the 1993 Factbook, demographic estimates for some countries (mostly African) have explicitly taken into account the effects of the growing impact of the HIV/AIDS epidemic. These countries are currently: The Bahamas, Benin, Botswana, Brazil, Burkina Faso, Burma, Burundi, Cambodia, Cameroon, Central African Republic, Democratic Republic of the Congo, Republic of the Congo, Cote d'Ivoire, Ethiopia, Gabon, Ghana, Guyana, Haiti, Honduras, Kenya, Lesotho, Malawi, Mozambique, Namibia, Nigeria, Rwanda, South Africa, Swaziland, Tanzania, Thailand, Togo, Uganda, Zambia, and Zimbabwe."
Population growth rate	MNAR	"The average annual percent change in the population, resulting from a surplus (or deficit) of births over deaths and the balance of migrants entering and leaving a country. The rate may be positive or negative. The growth rate is a factor in determining how great a burden would be imposed on a country by the changing needs of its people for infrastructure (e.g., schools, hospitals, housing, roads), resources (e.g., food, water, electricity), and jobs. Rapid population growth can be seen as threatening by neighboring countries."
Public debt	MNAR	"This entry records the cumulative total of all government borrowings less repayments that are denominated in a country's home currency. Public debt should not be confused with external debt, which reflects the foreign currency liabilities of both the private and public sector and must be financed out of foreign exchange earnings."

Railways	MNAR	"This entry states the total route length of the railway network and of its component parts by gauge, which is the measure of the distance between the inner sides of the load-bearing rails. The four typical types of gauges are: broad, standard, narrow, and dual. Other gauges are listed in a note. Some 60% of the world's railways use the standard gauge of 1.4 m (4.7 ft). Gauges vary by country and sometimes within countries. The choice of gauge during initial construction was mainly in response to local conditions and the intent of the builder. Narrow-gauge railways were cheaper to build and could negotiate sharper curves, broad-gauge railways gave greater stability and permitted higher speeds. Standard-gauge railways were a compromise between narrow and broad gauges."
Real GDP (purchasing power parity)	MNAR	"This entry gives the gross domestic product (GDP) or value of all final goods and services produced within a nation in a given year. A nation's GDP at purchasing power parity (PPP) exchange rates is the sum value of all goods and services produced in the country valued at prices prevailing in the United States in the year noted. This is the measure most economists prefer when looking at per-capita welfare and when comparing living conditions or use of resources across countries. The measure is difficult to compute, as a US dollar value has to be assigned to all goods and services in the country regardless of whether these goods and services have a direct equivalent in the United States (for example, the value of an ox-cart or non-US military equipment); as a result, PPP estimates for some countries are based on a small and sometimes different set of goods and services. In addition, many countries do not formally participate in the World Bank's PPP project that calculates these measures, so the resulting GDP estimates for these countries may lack precision. For many developing countries, PPP-based GDP measures are multiples of the official exchange rate (OER) measure. The differences between the OER- and PPP-denominated GDP values for most of the wealthy industrialized countries are generally much smaller."
Real GDP growth rate	MNAR	"This entry gives a country's real GDP annual growth rate, adjusted for seasonal unemployment and inflation. A country's growth rate is year-over-year, and not compounded."

Real GDP per capita	MNAR	"This entry shows real GDP, divided by population as of 1 July for the same year."
Refined petroleum products - exports	MNAR	"This entry is the country's total exports of refined petroleum products, in barrels per day (bbl/day)."
Refined petroleum products - imports	MNAR	"This entry is the country's total imports of refined petroleum products, in barrels per day (bbl/day)."
Refined petroleum products - production	MNAR	"This entry is the country's total output of refined petroleum products, in barrels per day (bbl/day). The discrepancy between the amount of refined petroleum products produced and/or imported and the amount consumed and/or exported is due to the omission of stock changes, refinery gains, and other complicating factors."
Reserves of foreign exchange and gold	MNAR	"This entry gives the dollar value for the stock of all financial assets that are available to the central monetary authority for use in meeting a country's balance of payments needs as of the end-date of the period specified. This category includes not only foreign currency and gold, but also a country's holdings of Special Drawing Rights in the International Monetary Fund, and its reserve position in the Fund."
Revenue from coal	MNAR	"This entry refers to the economic profits, expressed as a percentage of a country's GDP, from the extraction of coal. These profits equal coal gross revenues minus cost(s) to extract the coal. Other sources may refer to this field as coal rents."
Revenue from forest resources	MNAR	"This entry refers to the economic profits, expressed as a percentage of a country's GDP, from the harvesting of forests (e.g., lumber and timber industries). These profits equal forest gross revenues minus costs to harvest the forest. Other sources may refer to this field as forest rents."
Roadways	MNAR	"This entry gives the total length of the road network and includes the length of the paved and unpaved portions."

Taxes and other revenues	MNAR	"This entry records total taxes and other revenues received by the national government during the time period indicated, expressed as a percent of GDP. Taxes include personal and corporate income taxes, value added taxes, excise taxes, and tariffs. Other revenues include social contributions - such as payments for social security and hospital insurance - grants, and net revenues from public enterprises. Normalizing the data, by dividing total revenues by GDP, enables easy comparisons across countries, and provides an average rate at which all income (GDP) is paid to the national level government for the supply of public goods and services."
Telephones - fixed lines	MNAR	"This entry gives the total number of fixed telephone lines in use, as well as the number of subscriptions per 100 inhabitants."
Telephones - mobile cellular	MNAR	"This entry gives the total number of mobile cellular telephone subscribers, as well as the number of subscriptions per 100 inhabitants. Note that because of the ubiquity of mobile phone use in developed countries, the number of subscriptions per 100 inhabitants can exceed 100."
Tobacco use	MAR	"This entry measures the age standardized prevalence of tobacco use, whether smoked or smokeless or both, among persons 15 years and older for the total population, and separately for the male and female populations."
Total fertility rate	MAR	"This entry gives a figure for the average number of children that would be born per woman if all women lived to the end of their childbearing years and bore children according to a given fertility rate at each age. The total fertility rate (TFR) is a more direct measure of the level of fertility than the crude birth rate, since it refers to births per woman. This indicator shows the potential for population change in the country. A rate of two children per woman is considered the replacement rate for a population, resulting in relative stability in terms of total numbers. Rates above two children indicate populations growing in size and whose median age is declining. Higher rates may also indicate difficulties for families, in some situations, to feed and educate their children and for women to enter the labor force. Rates below two children indicate populations decreasing in size and growing

		older. Global fertility rates are in general decline and this trend is most pronounced in industrialized countries, especially Western Europe, where populations are projected to decline dramatically over the next 50 years."
Unemployment rate	MAR	"This entry contains the percent of the labor force that is without jobs. Substantial underemployment might be noted."
Unemployment, youth ages 15-24	MAR	"This entry gives the percent of the total labor force ages 15-24 unemployed during a specified year."
Waterways	MNAR	"This entry gives the total length of navigable rivers, canals, and other inland bodies of water."

Appendix B: Clustering of Locations Python

```
from IPython.core.display import display, HTML
display(HTML("<style>.container { width:100% !important; }</style>"))
```

```
<IPython.core.display.HTML object>
```

```
#important libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from statsmodels.formula.api import ols
from scipy import stats
import statsmodels.api as sm
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
from sklearn_extra.cluster import KMedoids
import plotly.express as px
```

```
#setting default frame such that all results can be seen
pd.set_option('display.max_rows', 500)
pd.set_option('display.max_columns', 500)
pd.set_option('display.width', 1000)
```

```
pip install plotly
```

```
Requirement already satisfied: plotly in c:\users\alyss\anaconda3\anaconda\li
b\site-packages (5.11.0)
Requirement already satisfied: tenacity>=6.2.0 in c:\users\alyss\anaconda3\an
aconda\lib\site-packages (from plotly) (8.1.0)
Note: you may need to restart the kernel to use updated packages.
```

```
#importing clustering data: there are 257 locations with 55 different attribu
tes
df = pd.read_csv('Clustering Data.csv')
df.head()
```

Exploratory Data Analysis

```
## Univariate Analysis
```

```
df_desc = df.describe(include = 'all')
df_desc
```

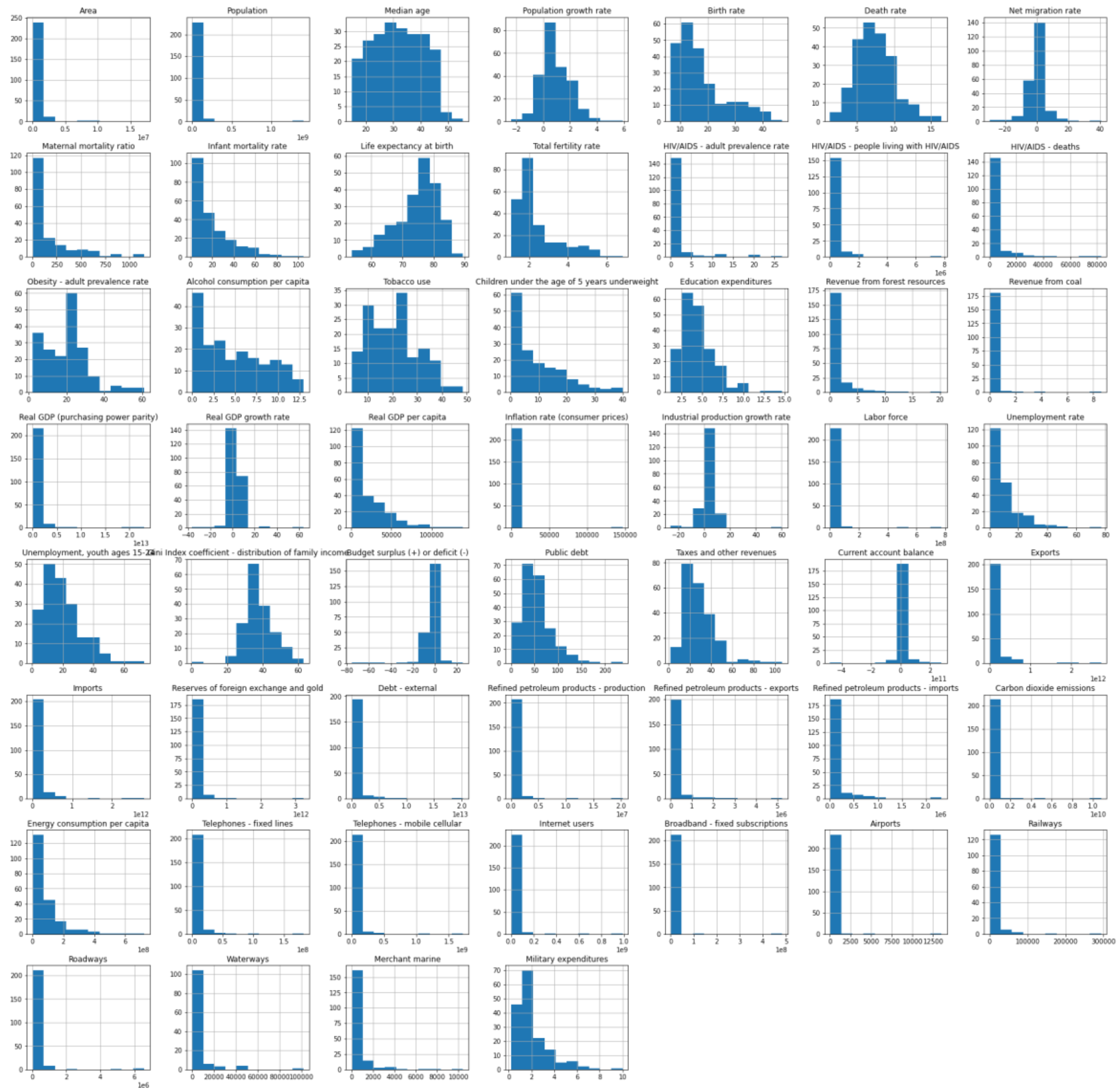
```
#features with the highest standard deviation
#we can see that in terms of wealth, there is the most distribution
std_sort = df_desc.sort_values(by = 'std', ascending = False, axis = 1)
std_sort.iloc[5][0:5]
```

```
Real GDP (purchasing power parity)       2.17078e+12
Debt - external                          1.75036e+12
Imports                                  2.93014e+11
Exports                                  2.91104e+11
Reserves of foreign exchange and gold    2.65623e+11
Name: std, dtype: object
```

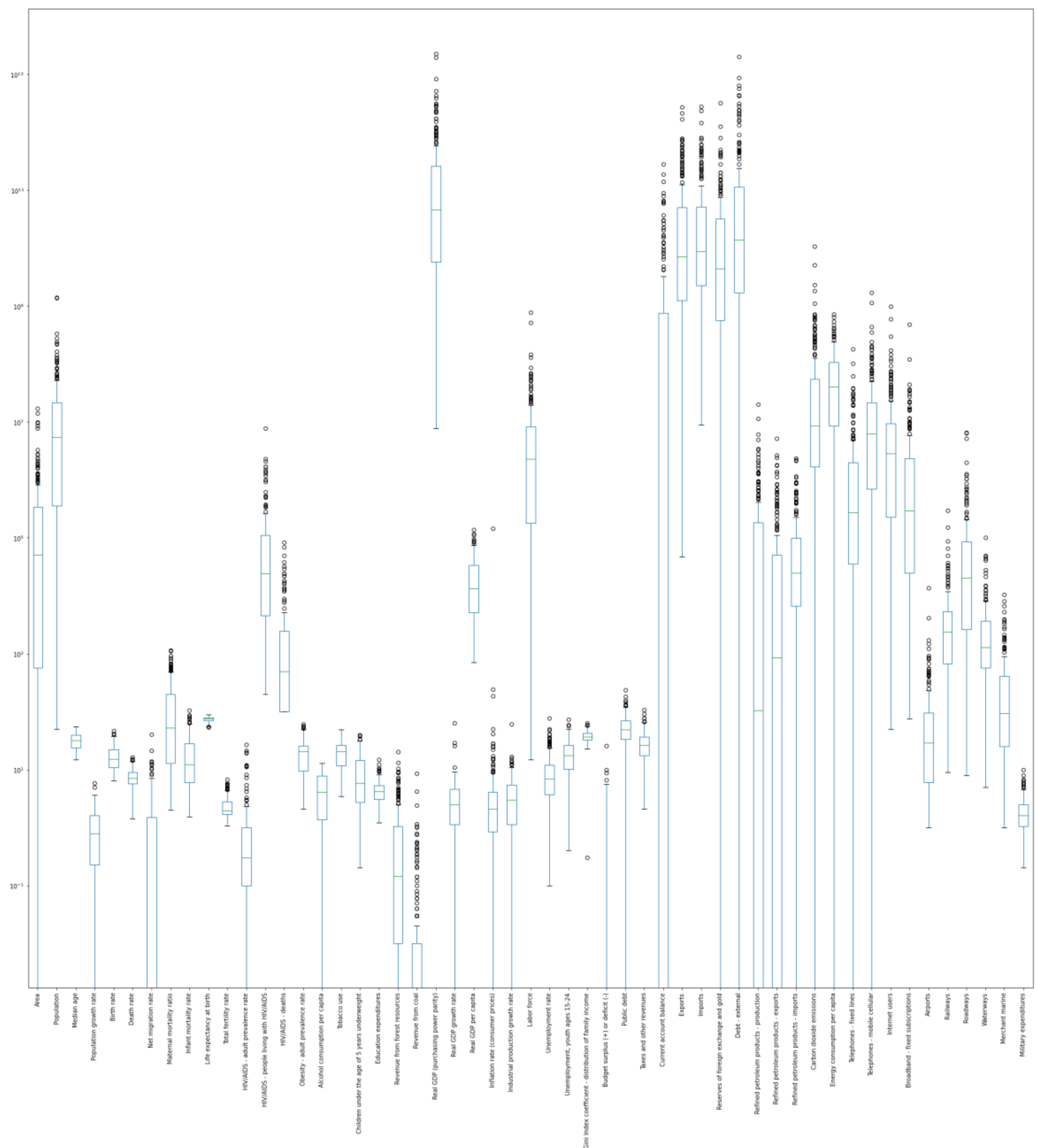
```
#features with the highest missing values
#some of these values are MNAR and assumed 0
miss_sort = df_desc.sort_values(by = 'count', ascending = True, axis = 1)
miss_sort.iloc[0][0:5]
```

```
Waterways                                        118
Railways                                         136
Children under the age of 5 years underweight    155
Tobacco use                                      164
HIV/AIDS - deaths                                167
Name: count, dtype: object
```

```
# we can see from our histograms that most data sets have some outliers
# besides in median age, growth rate, birth rate, death rate, infant mortalit
y rate, life expectancy, and fertility rate
# however, we can see some clear distributions in some of the data

df.hist(figsize = (30,30))
```



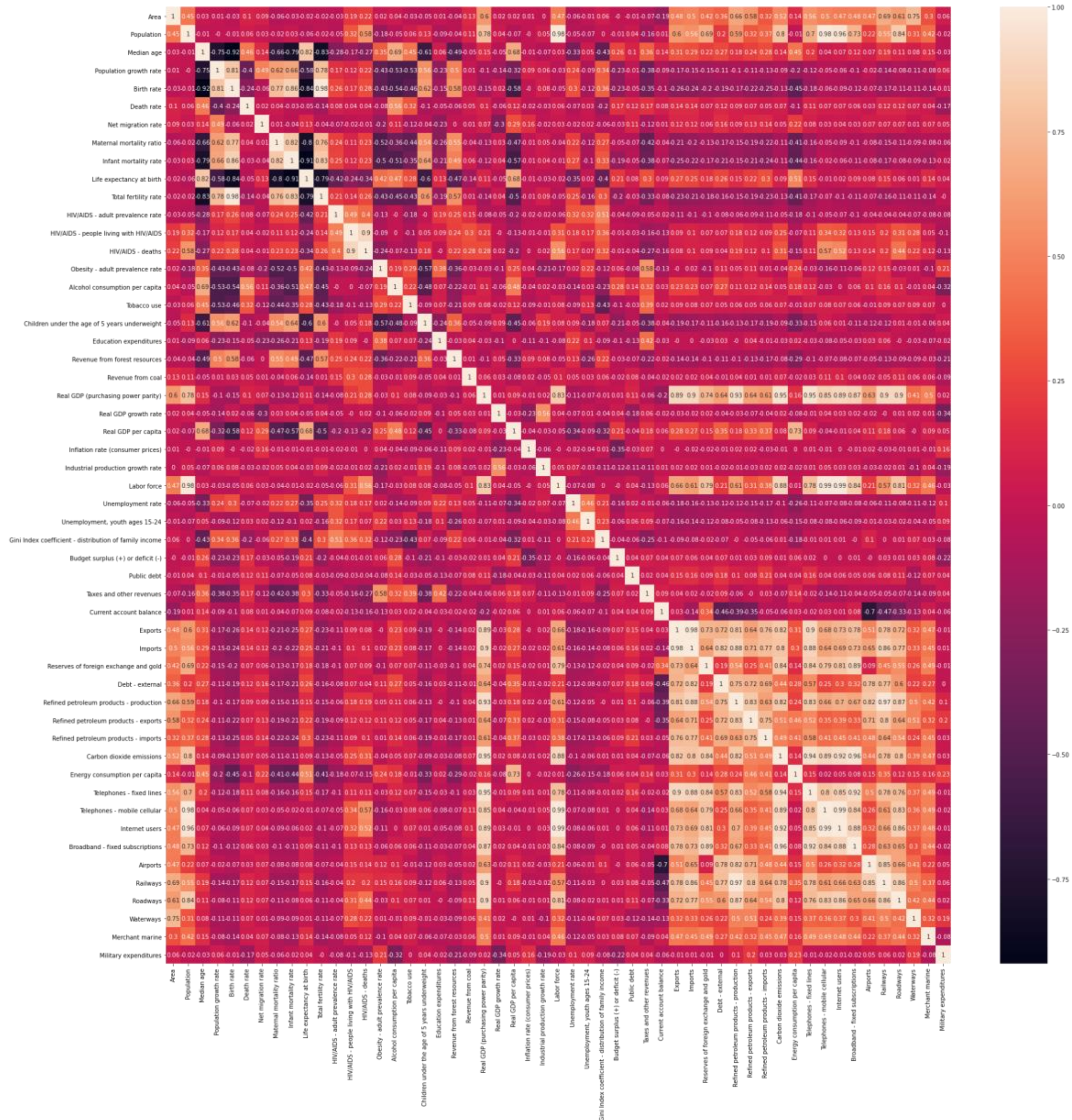

```
#with this box plot you can better visually see which variables consist of ou
tliers
#very few values have outliers
df.plot(kind = "box", figsize = (30,30), logy = True, rot = 90)
```




```
#Looking at the correlation between variables
#something interesting we see is that those "wealth" variables have high corr
elations
#some other notable high correlations are between population and labor force,
mobile telephone users, and internet users
#also high correlation between total fertility and birth rate
#these all ^^ make some intuitive sense
#the most unexplained high correlation is between labor force and telephone u
sers, and internet users

corr = df.corr()
```

```
plt.figure(figsize = (30, 30))
sns.heatmap(df.corr().round(2), annot = True)
```

```
<AxesSubplot:>
```



Data Preperation

```
#we know that the population values being 0 is meaningful; therefore, these v
alues will be removed because there are no inhabitants of these areas (i.e.,
no one to consider relocation for the refugee project)

df = df.drop(df[df['Population'].isnull()].index)
df.shape
```

```
(237, 56)

# The following attributes are considered to be missing not at random (MNAR)
and assumed to be 0 this was assumed with domain knowledge and an understandi
ng of the data collection process

df['Airports'] = df['Airports'].fillna(0)
df['Railways'] = df['Railways'].fillna(0)
df['Waterways'] = df['Waterways'].fillna(0)
df['Roadways'] = df['Roadways'].fillna(0)
df['Merchant marine'] = df['Merchant marine'].fillna(0)
df['Military expenditures'] = df['Military expenditures'].fillna(0)
df['Population growth rate'] = df['Population growth rate'].fillna(0)
df['Gini Index coefficient - distribution of family income'] = df['Gini Index
coefficient - distribution of family income'].fillna(0)
df['Revenue from coal'] = df['Revenue from coal'].fillna(0)
df['Reserves of foreign exchange and gold'] = df['Reserves of foreign exchang
e and gold'].fillna(0)
df['Industrial production growth rate'] = df['Industrial production growth ra
te'].fillna(0)
df['Education expenditures'] = df['Education expenditures'].fillna(0)
df['Current account balance'] = df['Current account balance'].fillna(0)
df['Debt - external'] = df['Debt - external'].fillna(0)
df['Revenue from forest resources'] = df['Revenue from forest resources'].fil
lna(0)
df['Public debt'] = df['Public debt'].fillna(0)
df['Energy consumption per capita'] = df['Energy consumption per capita'].fil
lna(0)
df['Refined petroleum products - imports'] = df['Refined petroleum products -
imports'].fillna(0)
df['Broadband - fixed subscriptions'] = df['Broadband - fixed subscriptions']
.fillna(0)
df['Refined petroleum products - production'] = df['Refined petroleum product
s - production'].fillna(0)
df['Refined petroleum products - exports'] = df['Refined petroleum products -
exports'].fillna(0)
df['Carbon dioxide emissions'] = df['Carbon dioxide emissions'].fillna(0)
df['Taxes and other revenues'] = df['Taxes and other revenues'].fillna(0)
df['Budget surplus (+) or deficit (-)'] = df['Budget surplus (+) or deficit (
-)'].fillna(0)
df['Real GDP growth rate'] = df['Real GDP growth rate'].fillna(0)
df['Telephones - mobile cellular'] = df['Telephones - mobile cellular'].filln
a(0)
df['Imports'] = df['Imports'].fillna(0)
df['Exports'] = df['Exports'].fillna(0)
df['Telephones - fixed lines'] = df['Telephones - fixed lines'].fillna(0)
df['Internet users'] = df['Internet users'].fillna(0)
df['Inflation rate (consumer prices)'] = df['Inflation rate (consumer prices)
'].fillna(0)
df['Real GDP (purchasing power parity)'] = df['Real GDP (purchasing power par
```

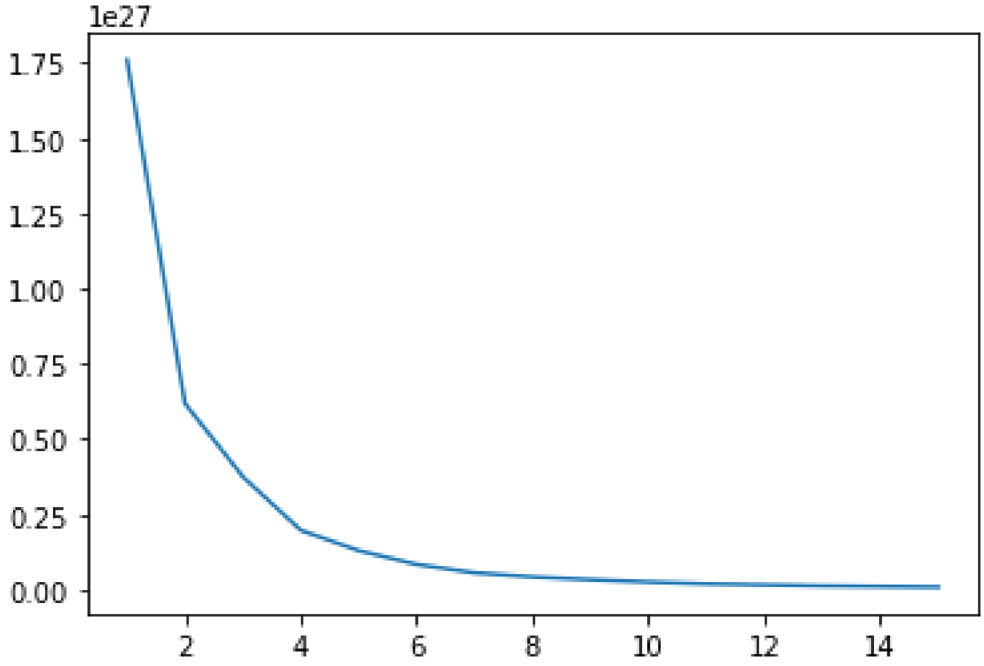
```
ity)'].fillna(0)
df['Real GDP per capita'] = df['Real GDP per capita'].fillna(0)
df['Net migration rate'] = df['Net migration rate'].fillna(0)

# #removing those observations that have more than 90% missing values (202 Re
maining) As they would not have enough information to cluster
# missing = (df.isnull().sum(axis=1) / 90) * 100
# df['missing'] = missing.tolist()
# df
# df.drop(df[df['missing'] >= 50].index, inplace = True)

# #getting rid of the new "missing" feature that was just made
# df = df.drop(['missing'], axis=1)

# looking at the co missing values for each attribute
# before we can do any feature engineering, we need to ensure all missing val
ues are filled. In order to avoid removing anymore data points such that the
data set is complete and no one is left out
df.isnull().sum().sort_values(ascending = False)
```

```
Children under the age of 5 years underweight            82
Tobacco use                                              73
HIV/AIDS - adult prevalence rate                         70
HIV/AIDS - people living with HIV/AIDS                   70
HIV/AIDS - deaths                                        70
Unemployment, youth ages 15-24                           53
Maternal mortality ratio                                 53
Alcohol consumption per capita                           48
Obesity - adult prevalence rate                          45
Unemployment rate                                        18
Median age                                               10
Infant mortality rate                                    10
Life expectancy at birth                                 10
Total fertility rate                                     10
Birth rate                                                9
Death rate                                                8
Labor force                                               6
ISO                                                       4
Real GDP per capita                                       0
Net migration rate                                        0
Population growth rate                                    0
Education expenditures                                    0
Revenue from forest resources                             0
Revenue from coal                                         0
Real GDP (purchasing power parity)                        0
Population                                                0
Area                                                      0
Region                                                    0
Real GDP growth rate                                      0
Industrial production growth rate                         0
```

```
Inflation rate (consumer prices)                        0
Military expenditures                                   0
Merchant marine                                         0
Waterways                                               0
Roadways                                                0
Railways                                                0
Airports                                                0
Broadband - fixed subscriptions                         0
Internet users                                          0
Telephones - mobile cellular                            0
Telephones - fixed lines                                0
Energy consumption per capita                           0
Carbon dioxide emissions                                0
Refined petroleum products - imports                    0
Refined petroleum products - exports                    0
Refined petroleum products - production                 0
Debt - external                                         0
Reserves of foreign exchange and gold                   0
Imports                                                 0
Exports                                                 0
Current account balance                                 0
Taxes and other revenues                                0
Public debt                                             0
Budget surplus (+) or deficit (-)                       0
Gini Index coefficient - distribution of family income  0
Location                                                0
dtype: int64
```

```
# # eliminating some redundant attributes based on correlation (focusing main
ly on problem areas i.e., many missing values)

# # based on correlation, we can see that the following are strongly correlat
ed and we will remove the variable with more missing values
# # birth rate and total fetility rate have a strong correlation as well as H
IV/AIDS - people living with HIV/AIDS and HIV/AIDS - deaths

# df = df.drop(['HIV/AIDS - deaths'], axis=1)
# df = df.drop(['Total fertility rate'], axis=1)

#a model based imputation was chosen for the missing values to keep the same
distribuitions and not create anymore
#filling the remaining variables with knn imputation

data = df.drop(['Region'], axis = 1)
data = data.drop(['Location'], axis = 1)
data = data.drop(['ISO'], axis = 1)

from sklearn.impute import KNNImputer
knn = KNNImputer(n_neighbors = 5, weights = 'uniform', metric = 'nan_euclidea
n')
```

```
nn = knn.fit_transform(data)
dataframe = pd.DataFrame(nn)
dataframe.columns = df.columns[2:55]
#scaling the variables

scaler = StandardScaler()
scaled_df = scaler.fit_transform(dataframe)

#checking the number of clusters necessary for the data

wcss = []

for i in range(1, 16):
    clustering = KMeans(n_clusters=i, init='k-means++', random_state=42)
    clustering.fit(dataframe)
    wcss.append(clustering.inertia_)

    ks = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15]
sns.lineplot(x = ks, y = wcss);
```




```
PAM = KMedoids(n_clusters=4).fit(scaled_df)

clusters = PAM.labels_
df['Cluster'] = clusters.tolist()
df['Cluster'].value_counts()

1    78
2    74
0    49
```

```
3    36
Name: Cluster, dtype: int64

print(df[['Location', 'Cluster', 'Region']].sort_values(ascending = False, by
= "Cluster"))

reg_clust = df.groupby(['Region', 'Cluster']).size()
reg_clust

Region                             Cluster
Africa                             0          43
                                   1           1
                                   2           9
                                   3           2
Australia and Oceania              0           1
                                   1           3
                                   2          18
                                   3           4
Central America and the Caribbean  1          15
                                   2          11
                                   3           6
Central Asia                       1           1
                                   2           3
                                   3           2
East and Southeast Asia            0           1
                                   1           7
                                   2           3
                                   3          10
Europe                             1          45
                                   2           4
                                   3           2
Middle East                        0           2
                                   1           1
                                   2          15
                                   3           1
North America                      1           4
                                   2           2
South America                      1           1
                                   2           7
                                   3           5
South Asia                         0           2
                                   2           2
                                   3           4
dtype: int64

gdp = df.groupby(['Cluster'])['Real GDP (purchasing power parity)'].agg('medi
an')
gdp

Cluster
0    3.205000e+10
```

```
1    1.165000e+11
2    2.656000e+10
3    5.384000e+10
Name: Real GDP (purchasing power parity), dtype: float64
```

```
ed = df.groupby(['Cluster'])['Education expenditures'].agg('median')
ed
```

```
Cluster
0    3.40
1    4.05
2    4.65
3    3.00
Name: Education expenditures, dtype: float64
```

```
exp = df.groupby(['Cluster'])['Life expectancy at birth'].agg('median')
exp
```

```
Cluster
0    65.300
1    80.585
2    75.750
3    73.080
Name: Life expectancy at birth, dtype: float64
```

```
co2 = df.groupby(['Cluster'])['Carbon dioxide emissions'].agg('median')
co2
```

```
Cluster
0     4041000.0
1    25752500.0
2     6159500.0
3     8266000.0
Name: Carbon dioxide emissions, dtype: float64
```

```
map = px.choropleth(df, locations="ISO",
                    color="Cluster",
                    hover_name="Location",
                    )
map.show()
```

Appendix C: Application of Optimization Model

```
#gurobi
from gurobipy import *
import pandas as pd


low = pd.read_excel('Parameters.xlsx', sheet_name='demand low')
med = pd.read_excel('Parameters.xlsx', sheet_name='demand med')
high = pd.read_excel('Parameters.xlsx', sheet_name='demand high')
host = pd.read_excel('Parameters.xlsx', sheet_name='hosts')


#sets


#origin locations
I = {101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114,
115, 116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129}

#host clusters - values from clustering
J = {1000, 1001, 1002, 1003}

# time periods
T = {1, 2, 3, 4}

# demand scenarios
W = ["L","M","H"]


# parameters


# upper bound for capacity change within one period for all clusters
ub = 30000000

#cost of opening the host clusters
co = {}

co[1000] = 69220000
co[1001] = 74720000
co[1002] = 75900000
co[1003] = 65000000

#cost of expansion for host clusters
ce = {}
ce[1000] = 150000000
ce[1001] = 150000000
ce[1002] = 150000000
```

```
ce[1003] = 150000000


# unit cost of deviation from goal by one individual
cd = 4

#initial resettlement capacity of host locations
φ = {}
φ[1000,0] = 69220000
φ[1001,0] = 74720000
φ[1002,0] = 75900000
φ[1003,0] = 65000000

# scenario probabilities
p = {}
for w in W:
    p[w] = 1/len(W)

#relocation demand for origin at different time periods under different
demand scenarios
d = {}

low = low.set_index('I')
med = med.set_index('I')
high = high.set_index('I')

for i in I:
    for t in T:
        d[i, t, "L"] = low.at[i,t]
        d[i, t, "M"] = med.at[i,t]
        d[i, t, "H"] = high.at[i,t]


#initialize model
model = Model("Adding clustering to a multi objective optimization model")


#decision variables


# if the cluster is opened during a period
y = {}


#capacity expansion for the upcoming period for clusters at a time period
Δφ = {}

#number of individuals assigned /relocation flow from an origin to a host at
a certain time under a demand scenario
x = {}
```

```
# resettlement demand
γ = {}



for j in J:
    for t in T:
        Δφ[j,t] = model.addVar(vtype=GRB.CONTINUOUS, name="Expansion", lb =
0)
        φ[j,t] = model.addVar(vtype=GRB.CONTINUOUS, name="Capacity", lb=0.0)
        y[j,t] = model.addVar(vtype=GRB.BINARY, name = "to open or not")


for w in W:
    for i in I:
        for j in J:
            for t in T:
                x[i,j,t,w] = model.addVar(vtype=GRB.CONTINUOUS, name="Flow",
lb=0.0)

for w in W:
    for i in I:
            for t in T:
                γ[i,t,w] = model.addVar(vtype=GRB.CONTINUOUS, name="demand",
lb=0.0)


#constraints


#can only open during one time period
for j in J:
    model.addConstr(quicksum(y[j,t] for t in T) >= 1)


#making sure that the capacity of the previous period and the previous
periods expansion is equal to the current capacity
for j in J:
    for t in set(itertools.islice(T,1,len(T),1)):
        model.addConstr(φ[j,t] == φ[j,t-1] + Δφ[j,t-1])
        model.addConstr(Δφ[j,t-1] <= quicksum(y[j,m] for m in list(range(1,t-
1+1))) * ub)



#the initial capcaity is equal to the capacity in the first period
for j in J:
    model.addConstr(φ[j,0] == φ[j,1])

#the flow cannot be larger than the capacity during a time period beginning
at period 1
for i in I:
    for w in W:
```

```
        for j in J:
            for t in T:
                model.addConstr(quicksum(x[i,j,m,w] for i in I for m in 
list(range(1,t+1))) <= φ[j,t] * quicksum(y[j,m] for m in list(range(1,t+1))))


#all of the flow cannot be greater than or equal to the demand at the 
beginning of a period
for i in I:
    for w in W:
        for t in T:
            model.addConstr(quicksum(x[i,j,t,w] for j in J) <= γ[i,t,w])

#demand for the first time period is equal to the forcasted demand in the 
first time period
for i in I:
    for w in W:
        model.addConstr(γ[i,1,w] == d[i,1,w])

#demad for a time period = demand of previous - flow + forcasted
for i in I:
    for j in J:
        for w in W:
            for t in set(itertools.islice(T,1,len(T),1)):
                model.addConstr(γ[i,t,w] == γ[i,t-1,w] - quicksum(x[i,j,t-
1,w] for j in J) + d[i,t,w])
```

In [9]:

```
#objective
obj1 = quicksum(co[j] * y[j,t] for j in J for t in T)

obj2 = quicksum(ce[j] * Δφ[j,t] for j in J for t in T)

for w in W:
    obj3 = quicksum(p[w] for w in W) * ((cd) * (quicksum(γ[i,t,w] - 
x[i,j,t,w] for i in I for j in J for t in T)))

for w in W:
    for i in I:
        for j in J:
            obj4 = quicksum(p[w] for w in W) * quicksum((γ[i,t,w] - 
quicksum(x[i,j,t,w] for j in J)) for i in I)

obj = obj1 + obj2 + obj3 + obj4

model.setObjective(obj, GRB.MINIMIZE)

model.optimize()
Gurobi Optimizer version 9.5.2 build v9.5.2rc0 (win64)
Thread count: 4 physical cores, 8 logical processors, using up to 8 threads
Optimize a model with 1511 rows, 1788 columns and 8183 nonzeros
Model fingerprint: 0xe796ed29
Model has 1392 quadratic constraints
```

```
Variable types: 1772 continuous, 16 integer (16 binary)
Coefficient statistics:
  Matrix range     [1e+00, 3e+07]
  QMatrix range    [1e+00, 1e+00]
  QLMatrix range   [1e+00, 1e+00]
  Objective range  [4e+00, 2e+08]
  Bounds range     [1e+00, 1e+00]
  RHS range        [1e+00, 8e+07]
Presolve removed 1355 rows and 1189 columns
Presolve time: 0.02s
Presolved: 240 rows, 699 columns, 2118 nonzeros
Presolved model has 64 SOS constraint(s)
Variable types: 651 continuous, 48 integer (48 binary)
Found heuristic solution: objective 4.372220e+09

Root relaxation: objective 4.238000e+09, 306 iterations, 0.00 seconds (0.00 w
ork units)

    Nodes    |    Current Node    |     Objective Bounds      |     Work
 Expl Unexpl |  Obj  Depth IntInf | Incumbent    BestBd   Gap | It/Node Time

     0     0 4.2380e+09    0   23 4.3722e+09 4.2380e+09  3.07%     -    0s
     0     0 4.2380e+09    0   46 4.3722e+09 4.2380e+09  3.07%     -    0s
H    0     0                     4.313900e+09 4.2380e+09  1.76%     -    0s
     0     0 4.2380e+09    0   18 4.3139e+09 4.2380e+09  1.76%     -    0s
H    0     0                     4.312720e+09 4.2380e+09  1.73%     -    0s
H    0     0                     4.303000e+09 4.2380e+09  1.51%     -    0s
     0     0 4.2380e+09    0    4 4.3030e+09 4.2380e+09  1.51%     -    0s
     0     0 4.2380e+09    0    3 4.3030e+09 4.2380e+09  1.51%     -    0s
     0     0 4.2380e+09    0   15 4.3030e+09 4.2380e+09  1.51%     -    0s
     0     0 4.2380e+09    0   32 4.3030e+09 4.2380e+09  1.51%     -    0s
     0     0 4.2380e+09    0    2 4.3030e+09 4.2380e+09  1.51%     -    0s
     0     0     cutoff    0      4.3030e+09 4.3030e+09  0.00%     -    0s

Cutting planes:
  Implied bound: 26
  Flow cover: 5
  Relax-and-lift: 13

Explored 1 nodes (1452 simplex iterations) in 0.20 seconds (0.06 work units)
Thread count was 8 (of 8 available processors)

Solution count 4: 4.303e+09 4.31272e+09 4.3139e+09 4.37222e+09

Optimal solution found (tolerance 1.00e-04)
Best objective 4.303000000000e+09, best bound 4.303000000000e+09, gap 0.0000%
```

In [10]:

```
obj1.getValue()
```

Out[10]:

```
349840000.0
```

In [11]:

```
obj2.getValue()
```

Out[11]:

```
0.0
```

In [12]:

```
obj3.getValue()
```

Out[12]:

```
3953160000.0
```

In [13]:

```
obj4.getValue()
```

Out[13]:

```
0.0
```

Without Clusters

In [15]:

```
#sets


#origin locations
I = {101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114,
115, 116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129}

#host clusters - values from clustering
J = {1000, 1001, 1002, 1003, 1004, 1005, 1006, 1007, 1008, 1009, 1010, 1011}

# time periods
T = {1, 2, 3, 4}

# demand scenarios
W = ["L","M","H"]
```

In [16]:

```
# parameters


# upper bound for capacity change within one period for all locations
ub = 30000000

#cost of opening the host locations
co = {}

co[1000] = 6922000
co[1001] = 41532000
co[1002] = 20766000

co[1003] = 2504000
co[1004] = 36108000
co[1005] = 33604000

co[1006] = 35840000
co[1007] = 24955000
co[1008] = 15105000

co[1009] = 18632000
co[1010] = 6492000
co[1011] = 39876000

#cost of expansion
```

```
ce = {}
ce[1000] = 5000000
ce[1001] = 5000000
ce[1002] = 5000000
ce[1003] = 5000000
ce[1004] = 5000000
ce[1005] = 5000000
ce[1006] = 5000000
ce[1007] = 5000000
ce[1008] = 5000000
ce[1009] = 5000000
ce[1010] = 5000000
ce[1011] = 5000000

# unit cost of deviation from goal by one individual
cd = 4

#initial resettlement capacity of host locations
φ = {}
φ[1000,0] = 6922000
φ[1001,0] = 41532000
φ[1002,0] = 20766000
φ[1003,0] = 2504000
φ[1004,0] = 36108000
φ[1005,0] = 33604000
φ[1006,0] = 35840000
φ[1007,0] = 24955000
φ[1008,0] = 15105000
φ[1009,0] = 18632000
φ[1010,0] = 6492000
φ[1011,0] = 39876000

# scenario probabilities
p = {}
for w in W:
    p[w] = 1/len(W)

#relocation demand for origin at different time periods under different
demand scenarios
# d = {}

# low = low.set_index('I')
# med = med.set_index('I')
# high = high.set_index('I')

# for i in I:
#     for t in T:
#         d[i, t, "L"] = low.at[i,t]
#         d[i, t, "M"] = med.at[i,t]
#         d[i, t, "H"] = high.at[i,t]
```

In [17]:

```
#decision variables
```

```
# if the cluster is opened during a period
y = {}


#capacity expansion for the upcoming period for clusters at a time period
Δφ = {}

#number of individuals assigned /relocation flow from an origin to a host at
a certain time under a demand scenario
x = {}


# resettlement demand
γ = {}



for j in J:
    for t in T:
        Δφ[j,t] = model.addVar(vtype=GRB.CONTINUOUS, name="Expansion", lb =
0)
        φ[j,t] = model.addVar(vtype=GRB.CONTINUOUS, name="Capacity", lb=0.0)
        y[j,t] = model.addVar(vtype=GRB.BINARY, name = "to open or not")


for w in W:
    for i in I:
        for j in J:
            for t in T:
                x[i,j,t,w] = model.addVar(vtype=GRB.CONTINUOUS, name="Flow",
lb=0.0)

for w in W:
    for i in I:
            for t in T:
                γ[i,t,w] = model.addVar(vtype=GRB.CONTINUOUS, name="demand",
lb=0.0)
```

In [18]:

```
#constraints


#can only open during one time period
for j in J:
    model.addConstr(quicksum(y[j,t] for t in T) <= 1)


#making sure that the capacity of the previous period and the previous
periods expansion is equal to the current capacity
```

```
for j in J:
    for t in set(itertools.islice(T,1,len(T),1)):
        model.addConstr(φ[j,t] == φ[j,t-1] + Δφ[j,t-1])
        model.addConstr(Δφ[j,t-1] <= quicksum(y[j,m] for m in list(range(1,t-
1+1))) * ub)



#the initial capcaity is equal to the capacity in the first period
for j in J:
    model.addConstr(φ[j,0] == φ[j,1])

#the flow cannot be larger than the capacity during a time period beginning
at period 1
for i in I:
    for w in W:
        for j in J:
            for t in T:
                model.addConstr(quicksum(x[i,j,m,w] for i in I for m in
list(range(1,t+1))) <= φ[j,t] * quicksum(y[j,m] for m in list(range(1,t+1))))


#all of the flow cannot be greater than or equal to the demand at the
beginning of a period
for i in I:
    for w in W:
        for t in T:
            model.addConstr(quicksum(x[i,j,t,w] for j in J) <= γ[i,t,w])

#demand for the first time period is equal to the forcasted demand in the
first time period
for i in I:
    for w in W:
        model.addConstr(γ[i,1, w] == d[i,1,w])

#demad for a time period = demand of previous - flow + forcasted
for i in I:
    for j in J:
        for w in W:
            for t in set(itertools.islice(T,1,len(T),1)):
                model.addConstr(γ[i,t,w] == γ[i,t-1,w] - quicksum(x[i,j,t-
1,w] for j in J) + d[i,t,w])
```

In [19]:

```
#objective
obj1 = quicksum(co[j] * y[j,t] for j in J for t in T)

obj2 = quicksum(ce[j] * Δφ[j,t] for j in J for t in T)

for w in W:
    obj3 = quicksum(p[w] for w in W) * ((cd) * (quicksum(γ[i,t,w] -
x[i,j,t,w] for i in I for j in J for t in T)))
```

```
for w in W:
    for i in I:
        for j in J:
            obj4 = quicksum(p[w] for w in W) * quicksum((γ[i,t,w] -
quicksum(x[i,j,t,w] for j in J)) for i in I)

obj = obj1 + obj2 + obj3 + obj4

model.setObjective(obj, GRB.MINIMIZE)

model.optimize()
Gurobi Optimizer version 9.5.2 build v9.5.2rc0 (win64)
Thread count: 4 physical cores, 8 logical processors, using up to 8 threads
Optimize a model with 6685 rows, 6456 columns and 65101 nonzeros
Model fingerprint: 0xeeb7f39e
Model has 6960 quadratic constraints
Variable types: 6392 continuous, 64 integer (64 binary)
Coefficient statistics:
  Matrix range     [1e+00, 3e+07]
  QMatrix range    [1e+00, 1e+00]
  QLMatrix range   [1e+00, 1e+00]
  Objective range  [4e+00, 4e+07]
  Bounds range     [1e+00, 1e+00]
  RHS range        [1e+00, 8e+07]

MIP start from previous solve produced solution with objective 1.51789e+10 (0
.13s)
MIP start from previous solve produced solution with objective 1.5173e+10 (0.
21s)
MIP start from previous solve produced solution with objective 1.50958e+10 (0
.28s)
MIP start from previous solve produced solution with objective 1.50127e+10 (0
.29s)
Loaded MIP start from previous solve with objective 1.50127e+10
Processed MIP start in 1.13 seconds (0.93 work units)

Presolve removed 6449 rows and 4833 columns
Presolve time: 0.08s
Presolved: 488 rows, 1809 columns, 6006 nonzeros
Presolved model has 78 SOS constraint(s)
Variable types: 1722 continuous, 87 integer (87 binary)

Root relaxation: objective 1.464573e+10, 797 iterations, 0.01 seconds (0.00 w
ork units)

    Nodes    |    Current Node    |     Objective Bounds      |     Work
 Expl Unexpl |  Obj  Depth IntInf | Incumbent    BestBd   Gap | It/Node Time

     0     0 1.4648e+10    0   51 1.5013e+10 1.4648e+10  2.43%     -    1s
     0     0 1.4690e+10    0   59 1.5013e+10 1.4690e+10  2.15%     -    1s
     0     0 1.4690e+10    0   59 1.5013e+10 1.4690e+10  2.15%     -    1s
     0     0 1.4690e+10    0   32 1.5013e+10 1.4690e+10  2.15%     -    1s
     0     0 1.4710e+10    0   51 1.5013e+10 1.4710e+10  2.01%     -    1s
     0     0 1.4710e+10    0   52 1.5013e+10 1.4710e+10  2.01%     -    1s
     0     0 1.4731e+10    0   27 1.5013e+10 1.4731e+10  1.88%     -    1s
```

```
     0     0 1.4731e+10    0   26 1.5013e+10 1.4731e+10  1.88%     -    1s
     0     0 1.4731e+10    0   43 1.5013e+10 1.4731e+10  1.88%     -    1s
     0     0 1.4731e+10    0   50 1.5013e+10 1.4731e+10  1.88%     -    1s
     0     0 1.4731e+10    0   31 1.5013e+10 1.4731e+10  1.88%     -    1s
     0     0 1.4731e+10    0   36 1.5013e+10 1.4731e+10  1.88%     -    1s
     0     0 1.4731e+10    0   30 1.5013e+10 1.4731e+10  1.88%     -    1s
     0     0 1.4731e+10    0   30 1.5013e+10 1.4731e+10  1.88%     -    1s
     0     2 1.4731e+10    0   30 1.5013e+10 1.4731e+10  1.88%     -    1s

Cutting planes:
  Implied bound: 78
  Flow cover: 18
  Relax-and-lift: 28

Explored 1065 nodes (84024 simplex iterations) in 3.19 seconds (2.46 work uni
ts)
Thread count was 8 (of 8 available processors)

Solution count 5: 1.50127e+10 1.50127e+10 1.50958e+10 ... 1.51789e+10

Optimal solution found (tolerance 1.00e-04)
Best objective 1.501272600000e+10, best bound 1.501272600000e+10, gap 0.0000%
```

In [20]:

```
obj1.getValue()
```

Out[20]:

```
282336000.0
```

In [21]:

```
obj2.getValue()
```

Out[21]:

```
0.0
```

In [22]:

```
obj3.getValue()
```

Out[22]:

```
14683296000.0
```

In [23]:

```
obj4.getValue()
```

Out[23]:

```
47093999.99999924
```