Improving Quality of Life: Optimal Financial Aid Allocation

University of Toronto: Faculty of Applied Science and Engineering
Victoria Piroian

1. Introduction and Problem Definition

It has been 25 years since Africa's last economic recession, and the COVID-19 pandemic has caused poverty levels across the continent to reach an all-time high [1]. Although international aid is often seen as a solution, it can have detrimental effects without a clear sense of purpose. It is crucial to use data analytics to guide international aid models. As a foreign aid organization with a prescribed budget, the team aims to determine an optimal funding allocation to the continent of Africa to improve quality of life, measured by the human development index (HDI). HDI is calculated by taking the geometric average of life expectancy, education, and income per capita [2]. Since life expectancy and income per capita are difficult to directly impact through financial aid, the team aims to predict HDI using a new set of country characteristics which include features that can be directly improved through financial aid. The team aims to explore the following research questions:

1. What is the predicted HDI in Africa using features other than life expectancy, education, and income per capita? Which country characteristics impact HDI the most?
2. What is the optimal funding allocation to different resources with the objective of maximizing HDI?

2. Data

The team will be using four datasets to predict the quality of life and optimize financial aid allocation. First, the Life Expectancy set by the Global Health Observatory [3] contains the life expectancy per country as well as features such as infant and adult mortality rate, alcohol consumption, health expenditure, BMI, Hepatitis B (HepB), polio, HIV/AIDS and measles rates, and schooling. Second, a water and sanitation dataset from the Oxford Martin School [4] describes the percentage of clean drinking water per country. A third dataset from Our World in Data [5] introduces the human development index feature by country, which will be the target variable in the model. Finally, a dataset which maps countries and continents was used in the Exploratory Data Analysis (EDA) [6].

3. Exploratory Data Analysis

After collecting the necessary data to carry out the desired prediction and optimization methods, the team sought further insights and synthesized the data through exploratory data analysis (EDA).

3.1 Continent Selection

The team sought insights from the data to inform which continent to focus on. This decision would be based on both the average life expectancy as well as the continent with the highest proportion of developing countries.

Before this decision was made, the team combined various datasets and filled in missing data (see Section 3.2 for the methods used to fill missing data). For instance, the life expectancy and country/continent datasets were merged, allowing each country in the life expectancy table to be tracked by its corresponding continent. Next, the number of developing countries were extracted for each continent, using the pandas 'groupby' method, which revealed that the African continent had the most developing countries (100%). Next, the average life expectancy was calculated and graphed for each continent, as life expectancy is considered one of the strongest indicators of a country's health and therefore can inform which continent may have the highest need for aid [7]. As seen in Figure 1, Africa had the lowest average life expectancy (58.67 years). The team narrowed the project scope to Africa, as it had the highest percentage of developing countries and the lowest average life expectancy.

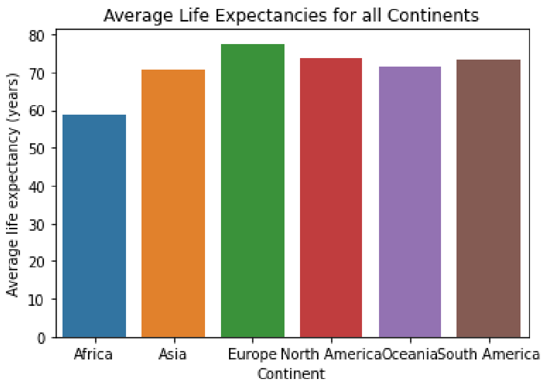


Figure 1: Life expectancy across all countries

To implement this decision, all the rows for non-African countries were dropped across all the datasets. A modular function to filter for African countries was created (see Appendix A), which was then applied to each of the datasets. The team noticed that there were variations in the naming of a few countries (i.e. “Ivory Coast” versus “Côte d’Ivoire”). In order to avoid issues while merging the datasets, the data was cleaned such that all countries with different spellings were renamed to ensure consistency.

3.2 Missing Data

After checking for missing values in the dataframe, the team developed a modular programming function for filling in missing data by country (see Appendix B). For columns with missing values that had some data for that country, the missing values were replaced with the average of the data for that country. For columns with missing values for all years for that country, the missing values were replaced with the average of all countries across all years in the dataset.

3.2 Data Merging

Upon replacing all missing data, the team was then able to merge the final three datasets into one. The three datasets were merged into a single data frame using a join on the country and year. Additionally, many of the columns were renamed more accordingly to better reflect the data they were describing. For instance, Hepatitis B was renamed Hepatitis B vaccines, and Measles was renamed Measles Cases to more easily differentiate between the number of people impacted by the disease versus the number of vaccinations available. This process can be seen in Appendix C.

3.3 Feature Selection

Across the three datasets, the team dropped any redundant, overlapping columns as well as columns which provided raw numbers instead of percentages. For example, the water and sanitation dataset provided both the percentage and the raw number of people per country with access to drinking water, and only the percentage column was kept.

Alcohol, schooling, and life expectancy were then dropped from the data. As previously mentioned, one of the motivations behind this report was to discover a way to predict HDI using a new set of features which could be directly impacted through financial aid. Accordingly, schooling and life expectancy, which are part of the existing HDI calculation, were dropped. Furthermore, alcohol was found to have a positive correlation with HDI. This is likely because alcohol is illegal in rougly 17% of African countries and so it is hypothesized that alcohol consumption would therefore be largely undereported in these developing countries [8]. Since a higher HDI is typically associated with developed countries (who have legalized alcohol consumption), the alcohol attribute likely isn’t directly impacting HDI. Instead, developed countries who have more lenient alcohol policies and are correctly reporting their alcohol consumption would already have higher HDI’s since they are developed.

3.4 Feature Engineering

To finish of the EDA on the merged dataset, the team performed feature engineering across attributes including the population, percentage expenditure, measles cases, and GDP. Firstly, the team plotted a histogram for every attribute in the final dataset to reveal which distributions were skewed and in need of feature engineering (see Figure 2 for before and after GDP was scaled). The team then selected the features listed above as needing featuring engineering as their distributions were skewed. The logarithm was first applied to each of these features and their distributions were plotted, however through trial and error, the team discovered that taking the squareroot of these features produced more normal looking distributions. Therefore, these original four features were dropped and were replaced by the squareroot of their values.

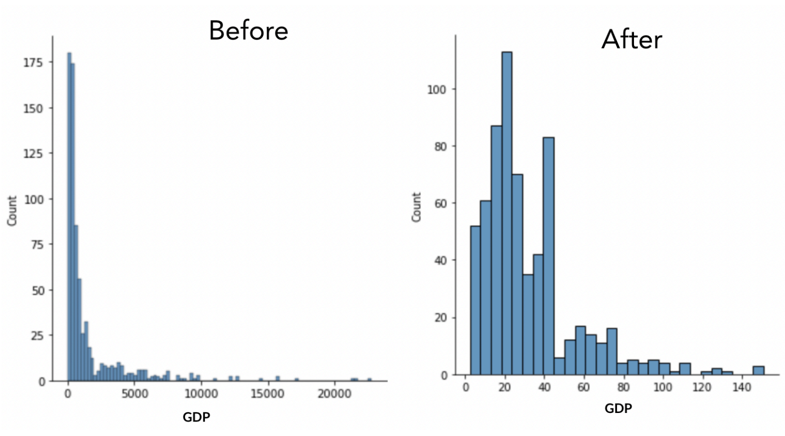


Figure 2: Impact of Feature Engineering on GDP

In addition to using domain knowledge, the team implemented a function to generate new features (see Appendix D). The function loops through each pair of features, multiplies them and then uses F-scores to evaluate if the product of these features is preferred over the individual features themselves. If that product is useful for predicting HDI, then its F-score would be above 1.5 with a p-value < 0.05 (implying the F-score is based on true data), and therefore the function would return that resulting feature is significant. However, after running the function on the current dataframe, the function indicated that no combination of features would return a new (and more useful) product. Therefore, besides feature scaling, no more feature engineering was performed.

Filling in missing data and conducting feature selection and feature engineering was completed after the data was split into training and testing sets in order to prevent the training set from becoming biased in favour of the testing set.

3. Prediction Model

After having conducted EDA, the team selected a method to predict the HDI across Africa. Since HDI is a continuous variable, a linear regression model was used:

$$y = \beta_0 + \beta_1 x_1 + \ \dots \ + \beta_k x_k$$

in which y represents HDI as the prediction target, $x_1,\dots, x_k$ represents all k features in the data, β_0 represents the y-intercept, and $\beta_1,\dots, \beta_k$ are the linear regression coefficients which indicate feature importance. The team aims to have a predicted $\widehat{y}$ HDI value that is close to the true average HDI of Africa.

3.1 Coding the Model and Research

In order to improve the accuracy and mitigate the bias when using a linear regression model, the team chose to implement k-fold cross validation, a resampling procedure which fits the model on a training dataset and evaluates it on the remaining test data set before discarding the model and repeating the procedure k times [9] in the model. Selecting an appropriate k value was crucial for the success of the model, due to the limited size of the dataset used. If k was too small there would not be sufficient training data to fit the model, and if k was too large there would not be sufficient data to evaluate the model [10]. In order to find a balance between these two issues, a K value of 5 was selected for the 816 rows of data which resulted in a 80-20 train-test split each time the resampling procedure ran.

In order to build the model, the k-fold and LinearRegression classes from the scikit-learn Python library were used along with the *Fit(), Predict()* and *Score()* methods. For each of the k=5 iterations, the training and testing sets had their missing data filled separately with the mean of the features for each nation. Feature engineering in the form of taking the square root for certain features was then completed in order for the data to follow a more normal distribution. Next, the model was fit with the training sets, the average of the predictions were recorded in a list, and the score of each iteration were recorded in a separate list. Finally, the linear regression coefficients were recorded for each iteration in order to determine the weight and importance of all the features in the model.

3.2 Prediction Results

The linear regression model resulted in R^2 scores of 0.77 and 0.67 on the training and testing datasets respectively. This suggests that there is minimal overfitting, as the data performs similarly on the training and testing sets. The average HDI for Africa was predicted to be 0.487, very close to the true average of 0.489 across the continent which was found using the HDI values in the data. This result was significant because it demonstrated that it was possible to calculate the HDI of countries using a series of other features that are not included in the conventional HDI equation with a relatively high accuracy.

The prediction model also informed the importance of various features when predicting HDI. This feature importance is indicated by the linear regression coefficients (β_i), shown in Figure 3 and are listed in Appendix E. It can be observed that some features have a positive coefficient value, and thus impact HDI positively, whereas some features have negative coefficient value and therefore have a detrimental effect on HDI. The features with the largest coefficient magnitudes have the greatest impact on HDI, such as access to drinking water, access to sanitation services, percentage expenditure, and access to handwashing facilities.

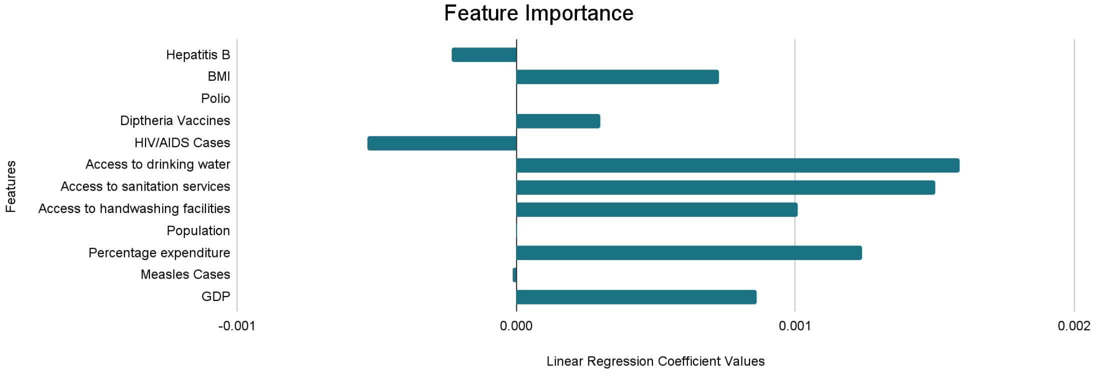


Figure 3: The linear regression coefficients (feature importance towards HDI) for each resource

4. Optimization Model

Next, an optimization model was created to determine the optimal funding allocation of the international aid organization's budget to different resources. The model parameters were tuned such that features which would be difficult to directly impact through financial aid – including BMI, GDP, population, and percentage expenditure – were dropped.

4.1 Coding the Model and Research

The optimization model is as follows:

$$max\ z = \sum_{i=1}^{8} \left|\beta_i\right| x_i \qquad (1)$$

$$s.t. \quad \sum_{i=1}^{8} c_i x_i \leq 1,300,000,000 \qquad (2)$$

$$x_i \leq D_i \qquad \forall\ i = 1,..., 8 \qquad (3)$$

$$\frac{c_i x_i}{1,300,000,000} \leq 0.15 \qquad \forall\ i = 1,..., 8 \qquad (4)$$

$$x_i \geq 0 \qquad \forall\ i = 1,..., 8 \qquad (5)$$

In this optimization model, the decision variables are the number of units of each resource that will be funded (x_i). The x values for disease features such as measles or HIV/AIDS now represent vaccine/treatment resources. One unit of a resource is defined to be the amount of that resource per capita. The objective function (1) is a sum of all x_i values multiplied by their linear regression coefficients, therefore, resources which have a greater impact on HDI are weighted more heavily. Several coefficients were negative, however since resources would be allocated towards their improvement, only the magnitude of the β_i values were of interest. Thus, the absolute values of the β_i values were taken. For example, $\beta_{HIV/AIDS\ Cases}$ was negative yet had a relatively high magnitude. Taking the absolute value ensured that its importance would be accounted for when determining how many treatment units should be funded. Several key constraints were implemented related to the total budget, the maximum demand of each resource, the maximum allowed funding allocation to each resource, and a non-negativity constraint on the decision variables.

Please refer to Appendix F for the coded model in Python. Due to the large magnitude of values in the model, the Python model was scaled down to have a total budget of 130 rather than 1.3B; large values in the model were multiplied by a factor of $\frac{130}{1,300,000,000}$.

4.1.1 Budgetary Constraint

The World Health Organization (WHO)'s budget allocated towards Africa for the 2022-2023 fiscal year is $1.3B USD [11]. This budget was implemented in the model through a constraint (2) which ensured that the sum of all decision variables (x_i) multiplied by their associated costs (c_i) should not exceed $1.3B USD. The costs for each resource were determined through secondary research, and are outlined below. Please see Appendix G for the research done to obtain each cost. All monetary values in the model are in USD. Although values that averaged all countries in Africa were sought, for some resources, data for just Sub-Saharan Africa was extrapolated to the entire continent. This is a limitation of the accuracy of the model. Additionally, it is important to note that although for some vaccinations, people must receive multiple doses in their lifetime, one dose is considered one resource unit since the scope of the model covers only one year (2022-2023) and no vaccines in the model have more than one required dose in a single year.

Table 1. Cost per capita of each resource

Resource	Cost per capita (c_i)
HepB vaccines	$0.2 [12]
Measles vaccines	$1.68 [13]
Polio vaccines	$0.78 [14]
Diphtheria vaccines	$77.5 [15]
HIV/AIDS treatment	$11 [16]
Basic drinking water	$3.0356 [17]
Basic sanitation services	$36.1174 [17]
Basic handwashing facilities	$0.17 [18]

4.1.2 Maximum Resource Demand Constraint

Next, a constraint was implemented which ensured that no more resource units would be allocated to a feature than what is needed (3). For example, based on how many people in Africa are in need of a certain resource i, an upper bound for x_i was determined. These upper bounds make up D_i, the demands for each resource, which are outlined in the table below. Please see Appendix H for the research based justifications of each demand value.

Table 2. Maximum demand of each resource (actual and scaled)

Resource	Actual Demand	Scaled Demand (D_i)
HepB vaccines	188,894,690 [19], [20]	18.889469
Measles vaccines	66,383,680 [19], [20]	6.683868
Polio vaccines	129,139,800 [19], [20]	12.91398
Diphtheria vaccines	124,835,140 [19], [20]	12.483514
HIV/AIDS treatment	25,700,000 [21]	2.57
Basic drinking water	418,000,000 [22]	41.8
Basic sanitation services	779,000,000 [22]	77.9

Basic handwashing facilities	1,086,692,400 [23]	108.66924

4.1.3 Maximum Funding Allocation Constraint

A statement made at the United Nations Economic and Social Council regarding funding and financing strategies states that realizing the 2030 Agenda requires “innovative financing to diversify the resource mobilization base” [24]. Since the impact of financial aid depends on many factors and can be difficult to predict, a constraint was implemented to ensure diversification of funding to different resources (4). A 15% cap on funding for a single resource was determined by the team, as it ensures that almost all features are part of the model without forcing every resource to be funded.

4.2 Optimization Results

From the optimization model, the optimal resource allocation was determined, as outlined in Table 3.

Table 3. Optimization model results

Resource	Scaled number of resource units	Actual number of resource units
HepB vaccines	18.889469	188,894,690
Measles vaccines	6.683868	66,838,680
Polio vaccines	12.913969	129,139,690
Diphtheria vaccines	0.251613	2,516,130
HIV/AIDS treatment	1.772727	17,727,270
Basic drinking water	6.423771	64,237,710
Basic sanitation services	0.539906	5,399,060
Basic handwashing facilities	108.66924	1,086,692,400

5. Discussion

The following section will describe the models’ strengths and weaknesses and future applications.

5.1 Optimal Funding Allocation

Based on the optimization model’s output for the number of resource units to invest in, the funding breakdown can be determined by multiplying the amount of each resource (x_i) by its cost (c_i). This optimal fund allocation breakdown is depicted in Figure 4. It can be seen that the vast majority of funding goes towards sanitation services, drinking water, HIV/AIDS treatments and Diphtheria vaccinations, which all met the 15% budget cap (constraint 4). Further analysis of binding constraints is discussed below.

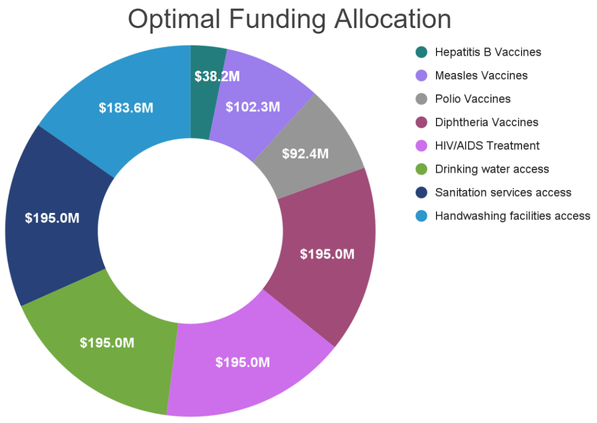


Figure 4: Optimal Predicted Fund Allocation

5.2 Sensitivity Analysis

The optimization model can be further investigated through sensitivity analysis (see Appendix I). By determining which constraints are binding and reach their limit, the team obtained a better understanding of how changes in the demand or allocation constraints will change the optimal funding allocation. Each of the eight features in the optimization model are bounded by constraints related to the maximum demand of the resource (constraint 3) and the maximum portion of the budget that can be allocated to it (constraint 4). As seen in Table 4, the HepB vaccines, measles cases, and access to basic handwashing facilities features have binding demand constraints and non-binding allocation constraints. This means that the total need for the units of these features was met without exceeding their respective allocated budget. If the allocation constraint of 15% of the budget was to be increased, the number of units for these features in the optimization results would not change. The Diptheria vaccines, HIV/AIDS cases, access to drinking water, and access to basic sanitation services have non-binding demand constraints and binding allocation constraints signifying that their respective allocated budget yet not all the demand for these features has been met. If the allocation constraint of 15% of the budget was to be increased, the number of units selected for these features would increase. The last feature left to analyse is polio which had both non-binding demand and allocation constraints. This suggests that this was the last feature to be invested in by the model, and at this time there was not enough budget to fill the need or reach its budget allocation constraint.

Table 4. Sensitivity Analysis Report

Resource	Demand constraint (3)	Maximum funding allocation constraint (4)
HepB vaccines	Binding	Not binding
Measles vaccines	Binding	Not binding
Polio vaccines	Not binding	Binding
Diphtheria vaccines	Not binding	Binding
HIV/AIDS treatment	Not binding	Binding
Basic drinking water	Not binding	Binding
Basic sanitation services	Not binding	Binding
Basic handwashing facilities	Binding	Not binding

5.2 Strengths and Weaknesses

An exploration of the strengths and weaknesses of this model is necessary in order to identify which parts can be applied to future application, and which need to be improved or reworked to obtain a more accurate and meaningful result. Firstly, the prediction model successfully provided an alternative way to evaluate HDI using features not included in the conventional calculation. This gives a greater variety of HDI prediction methods in the case that some of the features used in the original equation are unavailable. This new prediction model allows for a more realistic fund and resource allocation as it takes into account a greater variety of features than the conventional method and returns a very similar outcome for HDI. Allowing for the inclusion of a greater diversity of features will widen the scope and allow for these features to more accurately reflect the health and socioeconomic factors.

Although there are many strengths to this model which are important to recognize, it is equally vital to outline the weaknesses of this model to enable iterative improvement of the team's prediction and optimization results. The first shortcoming of this model stems from the outdated data the team used. The four data sets used in this model contain information for several countries before the year 2015.

There is no data for the years after 2015, yet the budget for the model was based on the 2022 - 2023 fund allocations for Africa. This misalignment in dates creates a disconnect between the data being analysed, and may impact the final result. Furthermore, the data in the years preceding 2015 do not accurately represent the current situation in the African continent as the COVID-19 pandemic has caused many social factors to undergo drastic changes which may result in a vastly different model. In addition to these shortcomings, the model fails to implement other factors from the economic, social, and political sector. This lack of variety in the types of features which the model contains does not account for implementation and distribution of foreign aid in the real world.

6. Next Steps and Model Versatility

In order to address some of the weaknesses listed above and further improve the funding allocation proposal for the United Nations, the team would like to allocate each resource by country after the optimal number of units for each of the features is determined. For example, after discovering that 17.7 million HIV/AIDS vaccines are needed in the optimal resource allocation, the team would like to further optimize the allocation of these vaccines to different countries in Africa based on which countries need the greatest supply of that resource. To address this problem, the team would need further research into the differences in demand and cost of each resource by country (rather than the entire continent). The team would then implement the model again, once for each of the eight resources. This time, the budget would be the funding that was allocated towards that resource in the original model and the decision variables would be the number of units of that resource to allocate to each country. So in the case of the HIV/AIDS vaccinations, which was allocated $195,000,000 in the original model, this capital would now be the new total budget for the budgetary constraint, the demand and cost of HIV/AIDS vaccines for each of the 54 countries would be the D_i and c_i respectively. The linear regression coefficients (β_i) would also need to be calculated in the prediction model, and would represent that country need for that resource. The model would solve for x_i, the number of units of HIV/AIDS vaccines to allocate to country i. After repeating this process for each of the eight resources, the team would now be able to propose how each of them could be optimally allocated by country, rather than for the entire continent.

Whether it be a similar financial aid model in another geographic region, or applied to other sectors such as the capital restructuring of a business where the board of directors is looking to invest funds into the business to maximize profit while considering various features related to their internal value chain and business operations, the versatility of this model allows it to be applicable in any scenario involving the maximization of a target value under a budgetary constraint.

7. Works Cited

[1] “What if Africa stops receiving foreign aid?,” *European Union Institute for Security Studies*, Jan. 20, 2022. https://www.iss.europa.eu/content/what-if-africa-stops-receiving-foreign-aid [Accessed 26-Nov-2022].

[2] M. Roser, “Human Development Index (HDI),” *Our World in Data*, Jul. 25, 2014. https://ourworldindata.org/human-development-index#:~:text=The%20HDI%20is%20calculated%20as,and%20expected%20years%20of%20schooling). [Accessed 07-Dec-2022].

[3] KumarRajarshi, “Life expectancy (WHO),” *Kaggle*, 10-Feb-2018. [Online]. Available: https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who. [Accessed: 05-Oct-2022].

[4] “Water, sanitation and hygiene (WASH) data explorer,” *Our World in Data*. [Online]. Available: https://ourworldindata.org/explorers/water-and-sanitation?tab=table&facet=none&Resource=Drinking%2Bwater&Level%2Bof%2BAccess=Safely%2Bmanaged&Residence=Total&Relative%2Bto%2Bpopulation=Share%2Bof%2Bpopulation&country=IND~USA~KEN~OWID_WRL~BGD~ZAF~CHN. [Accessed: 05-Oct-2022].

[5] M. Roser, “Human Development Index (HDI),” Our World in Data, Jul. 25, 2014. https://ourworldindata.org/human-development-index [Accessed: 29-Oct-2022].

[6] dbouquin, “IS_608/NanosatDB_munging/Countries-Continents.csv at master · dbouquin/IS_608,” GitHub, May 14, 2016. https://github.com/dbouquin/IS_608/blame/master/NanosatDB_munging/Countries-Continents.csv [Accessed: 29-Oct-2022].

[7] Z. B. Irfan and A. Nehra, “Analysing the aid effectiveness on the living standard: A check-up on Southeast Asian countries,” Journal of Urban Management, vol. 5, no. 1, pp. 23–31, Jun. 2016, doi: 10.1016/j.jum.2016.07.001. [Accessed: 29-Oct-2022].

[8] C. Ferreira-Borges, M. B. Esser, S. Dias, T. Babor, and C. D. H. Parry, “Alcohol Control Policies in 46 African Countries: Opportunities for Improvement,” Alcohol and Alcoholism, vol. 50, no. 4, pp. 470–476, Apr. 2015, doi: 10.1093/alcalc/agv036. [Accessed: 29-Oct-2022].

[9] J. Brownlee, “A Gentle Introduction to k-fold Cross-Validation - MachineLearningMastery.com,” MachineLearningMastery.com, May 22, 2018. https://machinelearningmastery.com/k-fold-cross-validation/ (accessed Dec. 08, 2022).

[10] Ludvig Renbo Olsen, “Multiple-k: Picking the number of folds for cross-validation,” R-project.org, Nov. 19, 2022. https://cran.r-project.org/web/packages/cvms/vignettes/picking_the_number_of_folds_for_cross-validation.html (accessed Dec. 08, 2022).

[11] World Health Organization, “Programme budget 2024-2025.” World Health Organization, 25-Aug-2022. [Accessed: 29-Oct-2022].

[12] A. Gosset et al., “The Costs of Introducing the Hepatitis B Birth Dose Vaccine into the National Immunization Programme in Senegal (NéoVac Study),” Vaccines, vol. 9, no. 5, p. 521, May 2021, doi: 10.3390/vaccines9050521. [Accessed: 29-Nov-2022].

[13] J. Brew and C. Sauboin, “A Systematic Review of the Incremental Costs of Implementing a New Vaccine in the Expanded Program of Immunization in Sub-Saharan Africa,” MDM Policy & Practice, vol. 4, no. 2, p. 238146831989454, Jul. 2019, doi: 10.1177/2381468319894546. [Accessed: 29-Nov-2022].

[14] "GPEI-Availability and price of inactivated polio vaccine," Polioeradication.org, 2022. https://polioeradication.org/news-post/availability-and-price-of-inactivated-polio-vaccine/#:~:text=The%20vaccine%20will%20now%20be [Accessed: 29-Nov-2022].

[15] "How Much Does Tetanus Diphtheria Vaccine Cost? - CostHelper," CostHelper, 2022. https://health.costhelper.com/td-vaccine.html [Accessed: 29-Nov-2022].

[16] A. Creese, K. Floyd, A. Alban, and L. Guinness, "Cost-effectiveness of HIV/AIDS interventions in Africa: a systematic review of the evidence," The Lancet, vol. 359, no. 9318, pp. 1635–1642, May 2002, doi: 10.1016/s0140-6736(02)08595-1. [Accessed: 29-Nov-2022].

[17] "Global costs and benefits of drinking-water supply and sanitation interventions to reach the MDG target and universal coverage." [Online]. Available: https://apps.who.int/iris/bitstream/handle/10665/75140/WHO_HSE_WSH_12.01_eng.pdf [Accessed: 29-Nov-2022].

[18] M. Freedman, S. D. Bennett, R. Rainey, R. Otieno, and R. Quick, "Cost analysis of the implementation of portable handwashing and drinking water stations in rural Kenyan health facilities," Journal of Water, Sanitation and Hygiene for Development, vol. 7, no. 4, pp. 659–664, Oct. 2017, doi: 10.2166/washdev.2017.010. [Accessed: 29-Nov-2022].

[19] S. Vanderslott, Saloni Dattani, F. Spooner, and M. Roser, "Vaccination," Our World in Data, May 10, 2013. https://ourworldindata.org/vaccination [Accessed: 29-Nov-2022].

[20] "Africa: population by age group 2021 | Statista," Statista, 2021. https://www.statista.com/statistics/1226211/population-of-africa-by-age-group/ (accessed Dec. 07, 2022). [Accessed: 29-Nov-2022].

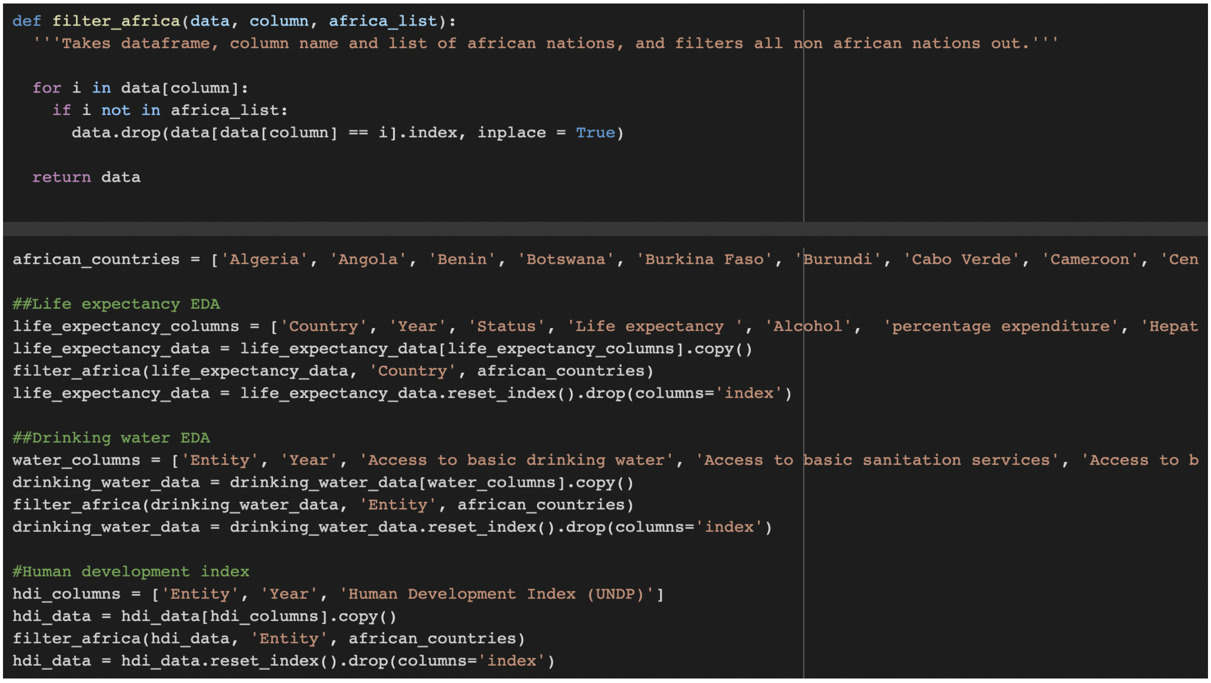
[21] "HIV/AIDS," WHO | Regional Office for Africa, 2020. https://www.afro.who.int/health-topics/hivaids#:~:text=The%20WHO%20African%20Region%20is,HIV%20in%20the%20African%20Region. [Accessed: 29-Nov-2022].

[22] "Africa to drastically accelerate progress on water, sanitation and hygiene – report," Unicef.org, 2022. https://www.unicef.org/senegal/en/press-releases/africa-drastically-accelerate-progress-water-sanitation-and-hygiene-report#:~:text=On%20the%20continent%2C%20however%2C%20418,still%20lack%20basic%20hygiene%20services [Accessed: 29-Nov-2022].

[23] "People with basic handwashing facilities including soap and water (% of population) - Sub-Saharan Africa | Data," Worldbank.org, 2022. https://data.worldbank.org/indicator/SH.STA.HYGN.ZS?locations=ZG [Accessed: 29-Nov-2022].

[24] ECOSOC, " Rethinking the funding and financing strategies to deliver 2030 Agenda ." United Nations. [Accessed: 03-Dec-2022].

Appendix A

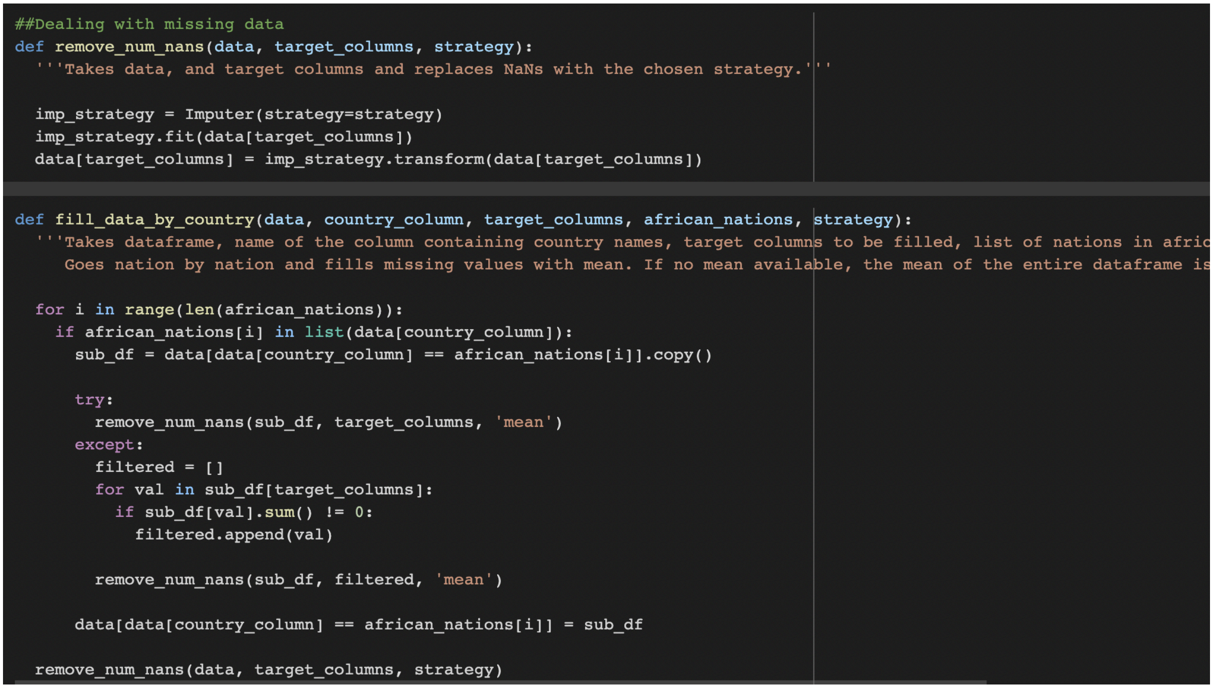


```
def filter_africa(data, column, africa_list):
  '''Takes dataframe, column name and list of african nations, and filters all non african nations out.'''

  for i in data[column]:
    if i not in africa_list:
      data.drop(data[data[column] == i].index, inplace = True)

  return data
```

```
african_countries = ['Algeria', 'Angola', 'Benin', 'Botswana', 'Burkina Faso', 'Burundi', 'Cabo Verde', 'Cameroon', 'Cen

##Life expectancy EDA
life_expectancy_columns = ['Country', 'Year', 'Status', 'Life expectancy ', 'Alcohol',  'percentage expenditure', 'Hepat
life_expectancy_data = life_expectancy_data[life_expectancy_columns].copy()
filter_africa(life_expectancy_data, 'Country', african_countries)
life_expectancy_data = life_expectancy_data.reset_index().drop(columns='index')

##Drinking water EDA
water_columns = ['Entity', 'Year', 'Access to basic drinking water', 'Access to basic sanitation services', 'Access to b
drinking_water_data = drinking_water_data[water_columns].copy()
filter_africa(drinking_water_data, 'Entity', african_countries)
drinking_water_data = drinking_water_data.reset_index().drop(columns='index')

#Human development index
hdi_columns = ['Entity', 'Year', 'Human Development Index (UNDP)']
hdi_data = hdi_data[hdi_columns].copy()
filter_africa(hdi_data, 'Entity', african_countries)
hdi_data = hdi_data.reset_index().drop(columns='index')
```

Figure 5. Function to Filter for African Countries used in EDA

Appendix B

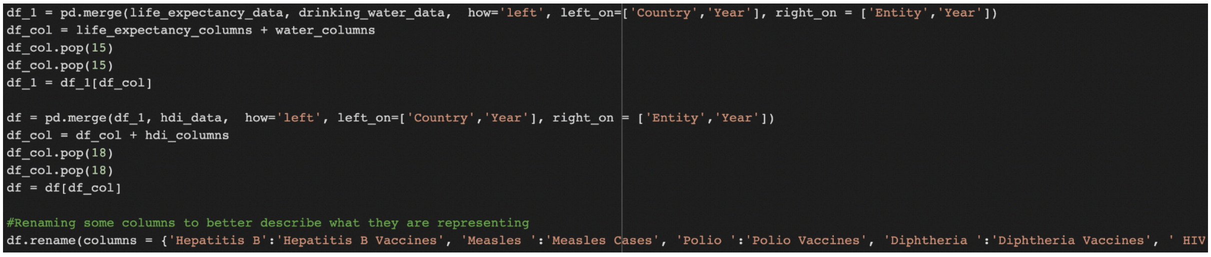


```
##Dealing with missing data
def remove_num_nans(data, target_columns, strategy):
  '''Takes data, and target columns and replaces NaNs with the chosen strategy.'''

  imp_strategy = Imputer(strategy=strategy)
  imp_strategy.fit(data[target_columns])
  data[target_columns] = imp_strategy.transform(data[target_columns])
```

```
def fill_data_by_country(data, country_column, target_columns, african_nations, strategy):
  '''Takes dataframe, name of the column containing country names, target columns to be filled, list of nations in afric
     Goes nation by nation and fills missing values with mean. If no mean available, the mean of the entire dataframe is

  for i in range(len(african_nations)):
    if african_nations[i] in list(data[country_column]):
      sub_df = data[data[country_column] == african_nations[i]].copy()

      try:
        remove_num_nans(sub_df, target_columns, 'mean')
      except:
        filtered = []
        for val in sub_df[target_columns]:
          if sub_df[val].sum() != 0:
            filtered.append(val)

        remove_num_nans(sub_df, filtered, 'mean')

      data[data[country_column] == african_nations[i]] = sub_df

  remove_num_nans(data, target_columns, strategy)
```

Figure 6. Missing Data Functions used in EDA

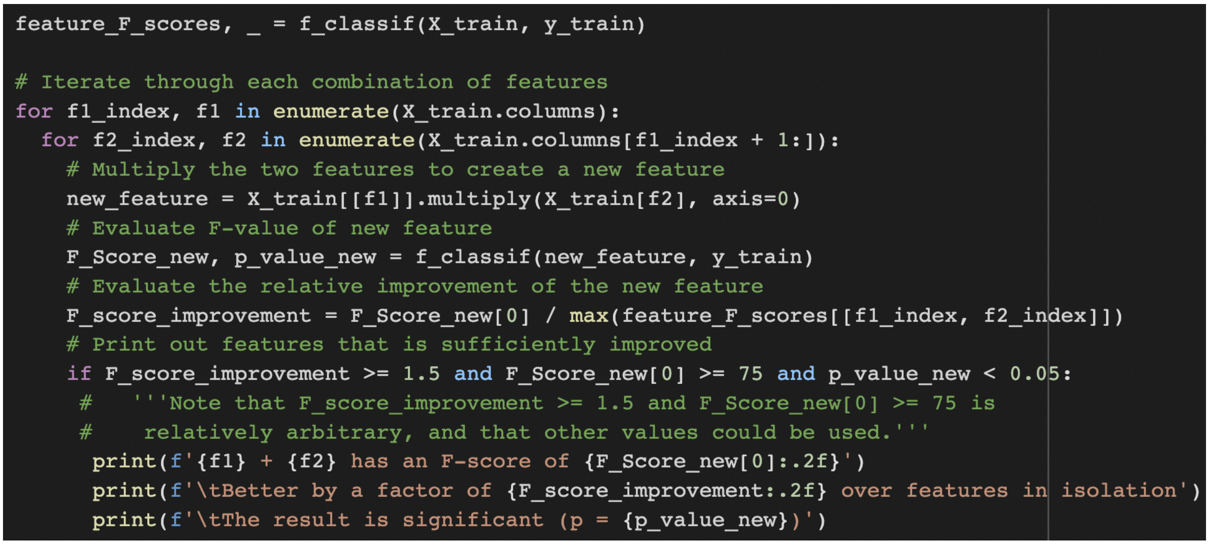
Appendix C



```
df_1 = pd.merge(life_expectancy_data, drinking_water_data,  how='left', left_on=['Country','Year'], right_on = ['Entity','Year'])
df_col = life_expectancy_columns + water_columns
df_col.pop(15)
df_col.pop(15)
df_1 = df_1[df_col]

df = pd.merge(df_1, hdi_data,  how='left', left_on=['Country','Year'], right_on = ['Entity','Year'])
df_col = df_col + hdi_columns
df_col.pop(18)
df_col.pop(18)
df = df[df_col]

#Renaming some columns to better describe what they are representing
df.rename(columns = {'Hepatitis B':'Hepatitis B Vaccines', 'Measles ':'Measles Cases', 'Polio ':'Polio Vaccines', 'Diphtheria ':'Diphtheria Vaccines', ' HIV
```

Figure 7. Merging the Datasets and Renaming Attributes in EDA

Appendix D



```
feature_F_scores, _ = f_classif(X_train, y_train)

# Iterate through each combination of features
for f1_index, f1 in enumerate(X_train.columns):
  for f2_index, f2 in enumerate(X_train.columns[f1_index + 1:]):
    # Multiply the two features to create a new feature
    new_feature = X_train[[f1]].multiply(X_train[f2], axis=0)
    # Evaluate F-value of new feature
    F_Score_new, p_value_new = f_classif(new_feature, y_train)
    # Evaluate the relative improvement of the new feature
    F_score_improvement = F_Score_new[0] / max(feature_F_scores[[f1_index, f2_index]])
    # Print out features that is sufficiently improved
    if F_score_improvement >= 1.5 and F_Score_new[0] >= 75 and p_value_new < 0.05:
     #   '''Note that F_score_improvement >= 1.5 and F_Score_new[0] >= 75 is
     #    relatively arbitrary, and that other values could be used.'''
      print(f'{f1} + {f2} has an F-score of {F_Score_new[0]:.2f}')
      print(f'\tBetter by a factor of {F_score_improvement:.2f} over features in isolation')
      print(f'\tThe result is significant (p = {p_value_new})')
```

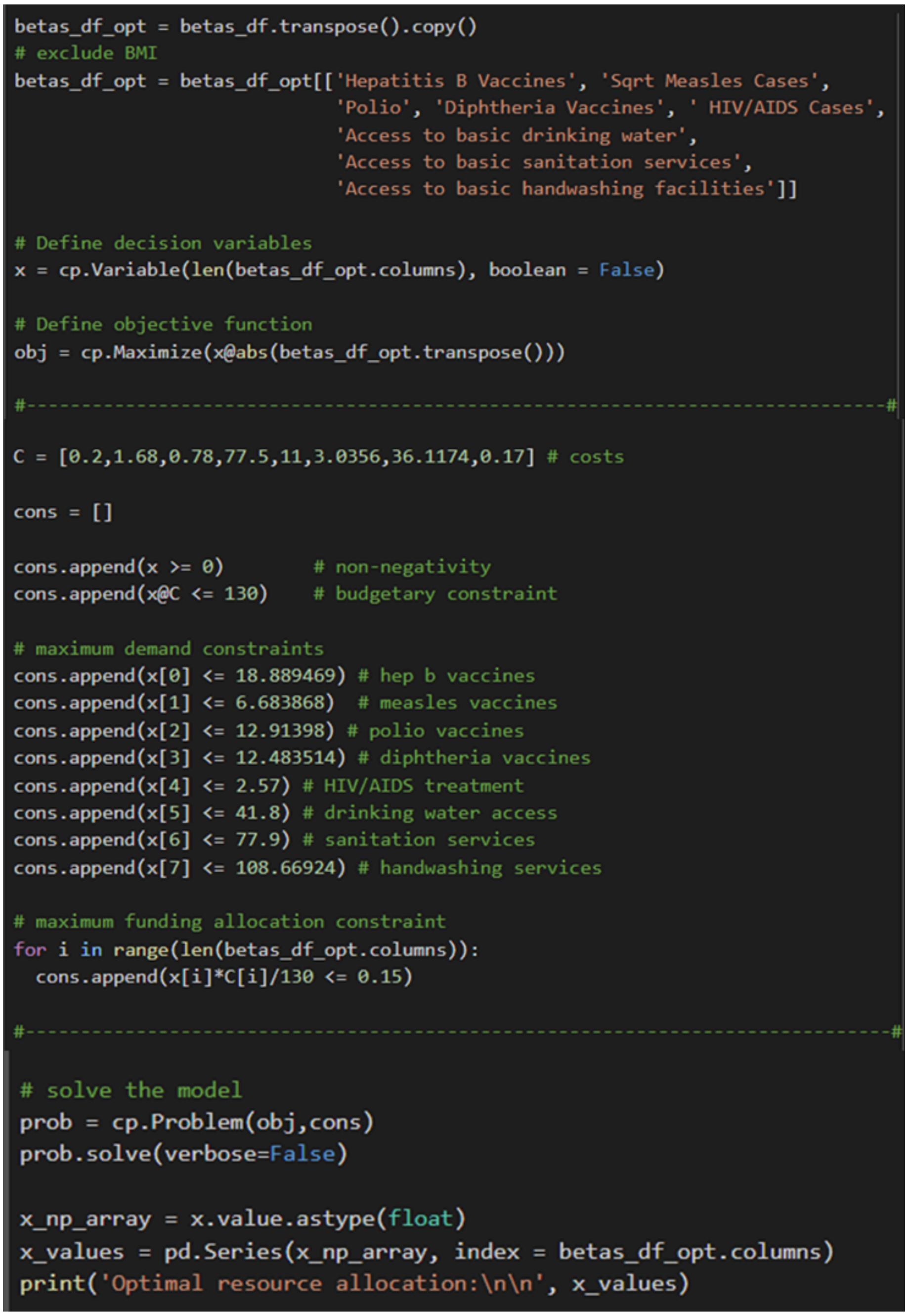
Figure 8. Identifying the need for feature engineering

Appendix E

Table 5. Linear regression coefficient values

Feature	Linear regression coefficient
Hepatitis B Vaccines	-0.000231
BMI	0.000727
Polio	-0.000000996
Diptheria Vaccines	0.000303
HIV/AIDS Cases	-0.000535
Access to drinking water	0.00159
Access to sanitation services	0.0015
Access to handwashing facilities	0.00101
Population	0.00000159
Percentage expenditure	0.00124
Measles Cases	-0.0000115
GDP	0.000861

Appendix F



```
betas_df_opt = betas_df.transpose().copy()
# exclude BMI
betas_df_opt = betas_df_opt[['Hepatitis B Vaccines', 'Sqrt Measles Cases',
                             'Polio', 'Diphtheria Vaccines', ' HIV/AIDS Cases',
                             'Access to basic drinking water',
                             'Access to basic sanitation services',
                             'Access to basic handwashing facilities']]

# Define decision variables
x = cp.Variable(len(betas_df_opt.columns), boolean = False)

# Define objective function
obj = cp.Maximize(x@abs(betas_df_opt.transpose()))

#--------------------------------------------------------------------------------#

C = [0.2,1.68,0.78,77.5,11,3.0356,36.1174,0.17] # costs

cons = []

cons.append(x >= 0)        # non-negativity
cons.append(x@C <= 130)    # budgetary constraint

# maximum demand constraints
cons.append(x[0] <= 18.889469) # hep b vaccines
cons.append(x[1] <= 6.683868)  # measles vaccines
cons.append(x[2] <= 12.91398) # polio vaccines
cons.append(x[3] <= 12.483514) # diphtheria vaccines
cons.append(x[4] <= 2.57) # HIV/AIDS treatment
cons.append(x[5] <= 41.8) # drinking water access
cons.append(x[6] <= 77.9) # sanitation services
cons.append(x[7] <= 108.66924) # handwashing services

# maximum funding allocation constraint
for i in range(len(betas_df_opt.columns)):
  cons.append(x[i]*C[i]/130 <= 0.15)

#--------------------------------------------------------------------------------#

# solve the model
prob = cp.Problem(obj,cons)
prob.solve(verbose=False)

x_np_array = x.value.astype(float)
x_values = pd.Series(x_np_array, index = betas_df_opt.columns)
print('Optimal resource allocation:\n\n', x_values)
```

Figure 9. Code for the optimization model

Appendix G

Table 6. Research for each resource's cost, used in the optimization model

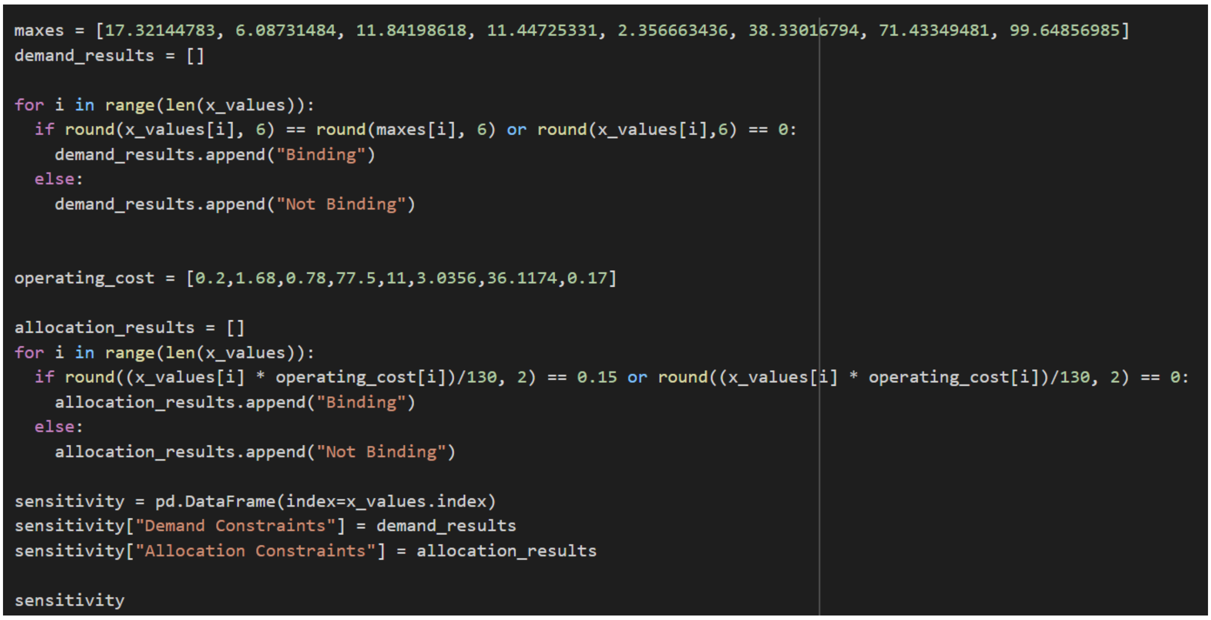
Resource	**Research**	**Cost per capita (c_i)**
HepB vaccines	A NeoVac study reports that the cost of vaccinating one person in Sub-Saharan Africa against HepB including pre-introduction and initial training costs was $0.2 in 2017 [12].	$0.2
Measles vaccines	A 2019 systematic review of the incremental costs of vaccine implementations in Sub-Saharan Africa reports that the mean cost per dose of Measles vaccines is $1.68 [13].	$1.68
Polio vaccines	According to the Polio Global Eradication Initiative, the cost of polio vaccines is $0.78 per dose for countries supported by the GAVI Alliance, which encompasses most African countries [14].	$0.78
Diphtheria vaccines	The cost of administering a diphtheria vaccine has an average cost of $42.5, and the cost of transporting a single vaccine is $35, making the overall cost for one Diphtheria vaccine $77.5 [15].	$77.5
HIV/AIDS treatment	A 2002 study shows that a case of HIV/AIDS can be prevented for a cost of $11 [16].	$11
Basic drinking water	Based on data from a 2012 WHO report, it was found that $4B and $1B were needed to provide drinking water to North Africa and Sub-Saharan Africa respectively [17]. Dividing these figures by the populations of North Africa and Sub-Saharan Africa and taking a weighted average based on the percentage of countries that each region covers in Africa, a final per capita cost of $3.0356 was found.	$3.0356
Basic sanitation services	According to the same 2012 WHO report used previously, $9.6B and $42B are required to provide sanitation services to North Africa and Sub-Saharan Africa respectively [17]. By the same process used for drinking water, the per capita cost of sanitation service provision was found to be $36.1174.	$36.1174
Basic handwashing facilities	A 2017 Kenyan study found that the cost of improved access to handwashing is $0.17 per individual [18].	$0.17

Appendix H

Table 7. Research for each resource's demand, used in the optimization model

Resource	Research	Demand (D_i)
HepB vaccines	HepB vaccine doses are administered to youth under 17 years old. 29% of youth need the HepB vaccine in Africa [19]. Based on the number of children under 17 in Africa [20], the demand is 188,894,690.	Actual: 188,894,690 Scaled: 18.889469
Measles vaccines	Measles vaccine doses are administered to newborns and children under 4 years old. 32% of children need the measles vaccine in Africa [19]. Based on the number of children under 4 in Africa [20], the demand is 66,383,680.	Actual: 66,383,680 Scaled: 6.683868
Polio vaccines	Polio vaccine doses are administered between birth and the age of 10. 30% of children need the polio vaccine in Africa [19]. Based on the number of children under 10 in Africa [20], the demand is 129,139,800.	Actual: 129,139,800 Scaled: 12.91398
Diphtheria vaccines	Diphtheria vaccine doses are administered to newborns and children under 10 years old. 29% of children need the diphtheria vaccine in Africa [19]. Based on the number of children under 10 in Africa [20], the demand is 124,835,140.	Actual: 124,835,140 Scaled: 12.483514
HIV/AIDS treatment	The WHO reports that 25.7 million people are living with HIV/AIDS in Africa [21].	Actual: 25,700,000 Scaled: 2.57
Basic drinking water	According to UNICEF, 418,000,000 people are in need of drinking water in Africa [22].	Actual: 418,000,000 Scaled: 41.8
Basic sanitation services	According to UNICEF, 779,000,000 people in Africa lack sanitation services [22].	Actual: 779,000,000 Scaled: 77.9
Basic handwashing facilities	The World Bank reports that 1,086,692,400 people are in need of handwashing facilities in Africa [23].	Actual: 1,086,692,400 Scaled: 108.66924

Appendix I



```
maxes = [17.32144783, 6.08731484, 11.84198618, 11.44725331, 2.356663436, 38.33016794, 71.43349481, 99.64856985]
demand_results = []

for i in range(len(x_values)):
  if round(x_values[i], 6) == round(maxes[i], 6) or round(x_values[i],6) == 0:
    demand_results.append("Binding")
  else:
    demand_results.append("Not Binding")


operating_cost = [0.2,1.68,0.78,77.5,11,3.0356,36.1174,0.17]

allocation_results = []
for i in range(len(x_values)):
  if round((x_values[i] * operating_cost[i])/130, 2) == 0.15 or round((x_values[i] * operating_cost[i])/130, 2) == 0:
    allocation_results.append("Binding")
  else:
    allocation_results.append("Not Binding")

sensitivity = pd.DataFrame(index=x_values.index)
sensitivity["Demand Constraints"] = demand_results
sensitivity["Allocation Constraints"] = allocation_results

sensitivity
```

Figure 10. Code to identify Binding and Non-Binding Constraints of the Optimization Model