## Analyzing Child Health Indicators to Optimize Global Healthcare Interventions

Zoha Ahmed

Data Analytics Report - 4000

ahmedz@rpi.edu

### Abstract

Child health remains a critical indicator of a nation's overall development and healthcare accessibility. This project investigates the distribution and relationships among child health intervention coverages and care-seeking behaviors using a dataset compiled from international health statistics. The dataset includes vaccination coverage rates, preventive health practices, and treatment-seeking behaviors across multiple countries and regions.

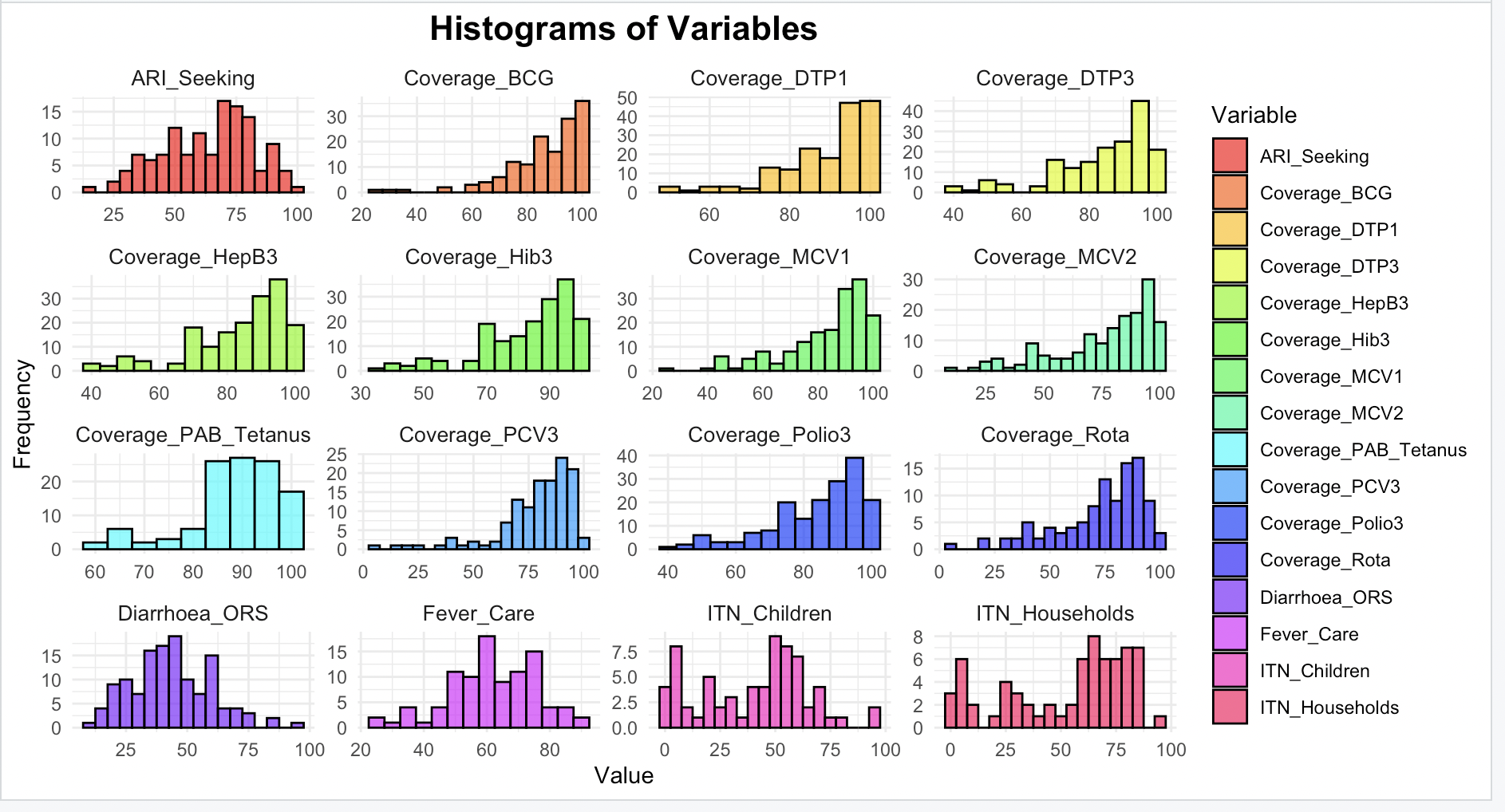
The study hypothesizes that higher immunization coverage correlates with improved care-seeking behaviors and reduced prevalence of preventable illnesses such as pneumonia and malaria. This hypothesis aligns with prior studies emphasizing the importance of comprehensive child health programs. By analyzing the relationships among these variables, this project aims to develop predictive models that can help identify factors contributing to low intervention coverage and care gaps.

The motivation for this project stems from its potential impact on global health policies, particularly in resource-limited settings. Previous work has highlighted the role of immunization and early treatment in reducing child mortality. However, this study aims to provide deeper insights into how these factors interact and influence outcomes, which may guide more effective resource allocation and intervention strategies.

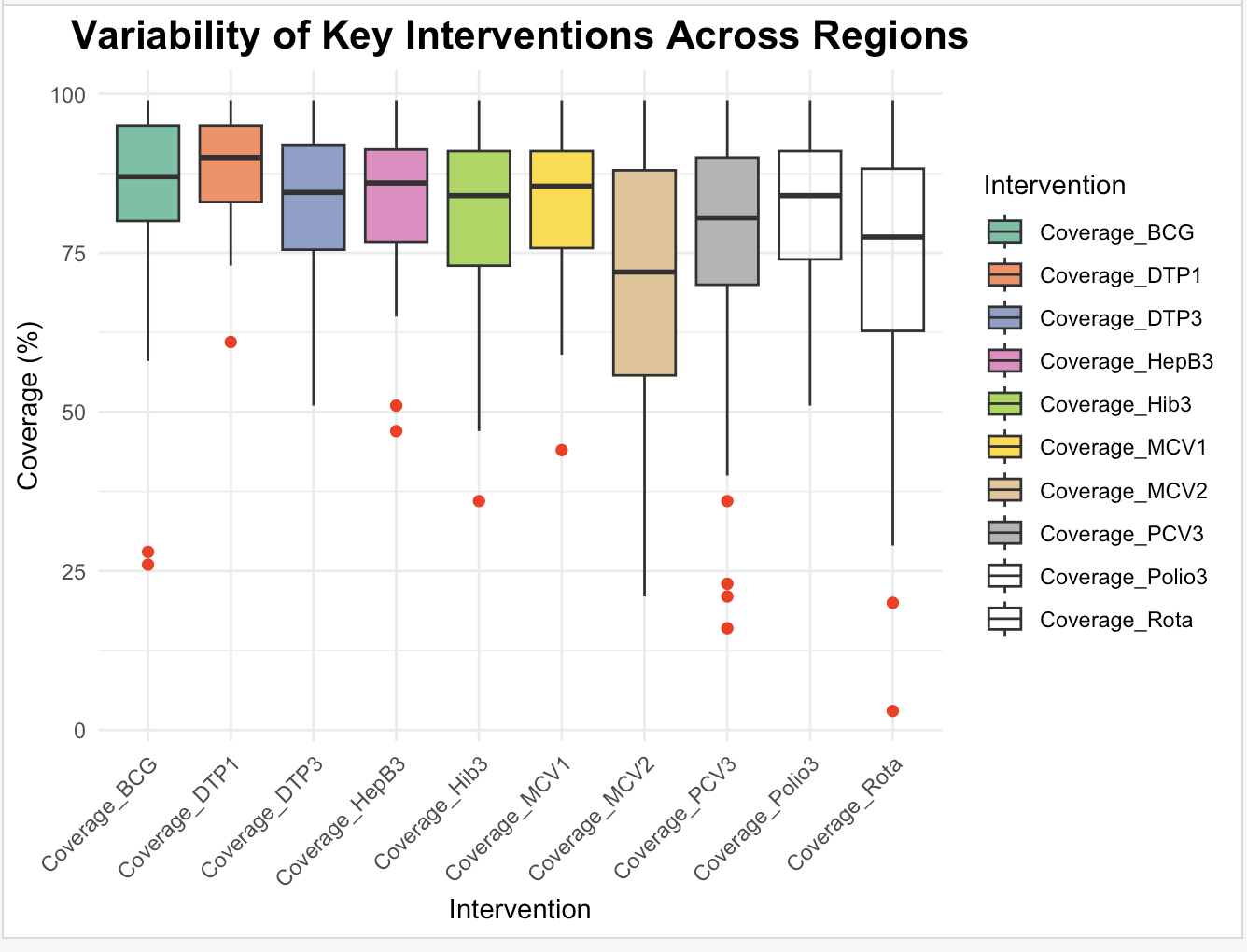
### Data Description and Preliminary Analysis

The dataset utilized in this project was selected based on its comprehensive coverage of child health indicators, including vaccination rates, care-seeking behaviors, and preventive health interventions. Sourced from UNICEF’s public data repository, the dataset is highly credible and well-documented, making it ideal for investigating global disparities in child health. The richness of the dataset, which combines intervention coverage and behavioral data, allowed for a deeper exploration of the relationships between these variables. This was particularly relevant to the project’s hypothesis, which seeks to understand how intervention coverage correlates with care-seeking behaviors to inform health policies.

Preliminary analyses began with descriptive statistics to explore the data’s central tendencies, variability, and range. The mean vaccination coverage for key interventions such as BCG, DTP1, and Polio3 ranged between 60% and 90%, reflecting substantial disparities across regions. Standard deviations for several variables indicated high variability, highlighting significant differences in healthcare access and intervention outcomes among countries. Additionally, behavioral metrics such as care-seeking for ARI (Acute Respiratory Infection) and fever exhibited relatively lower averages compared to vaccination coverage, pointing to gaps in healthcare utilization.



To visualize the data distributions, histograms were plotted for all key variables. These revealed important trends, including right-skewed distributions for vaccination coverage rates such as Coverage\_BCG and Coverage\_MCV1, where a significant number of countries reported high coverage levels. On the other hand, variables like ITN (Insecticide-Treated Nets) usage among children showed a broader distribution, with several countries performing poorly. Behavioral indicators such as Diarrhoea\_ORS (Oral Rehydration Solution) treatment had a more uniform distribution, with fewer extreme outliers, suggesting more consistent adoption across regions.



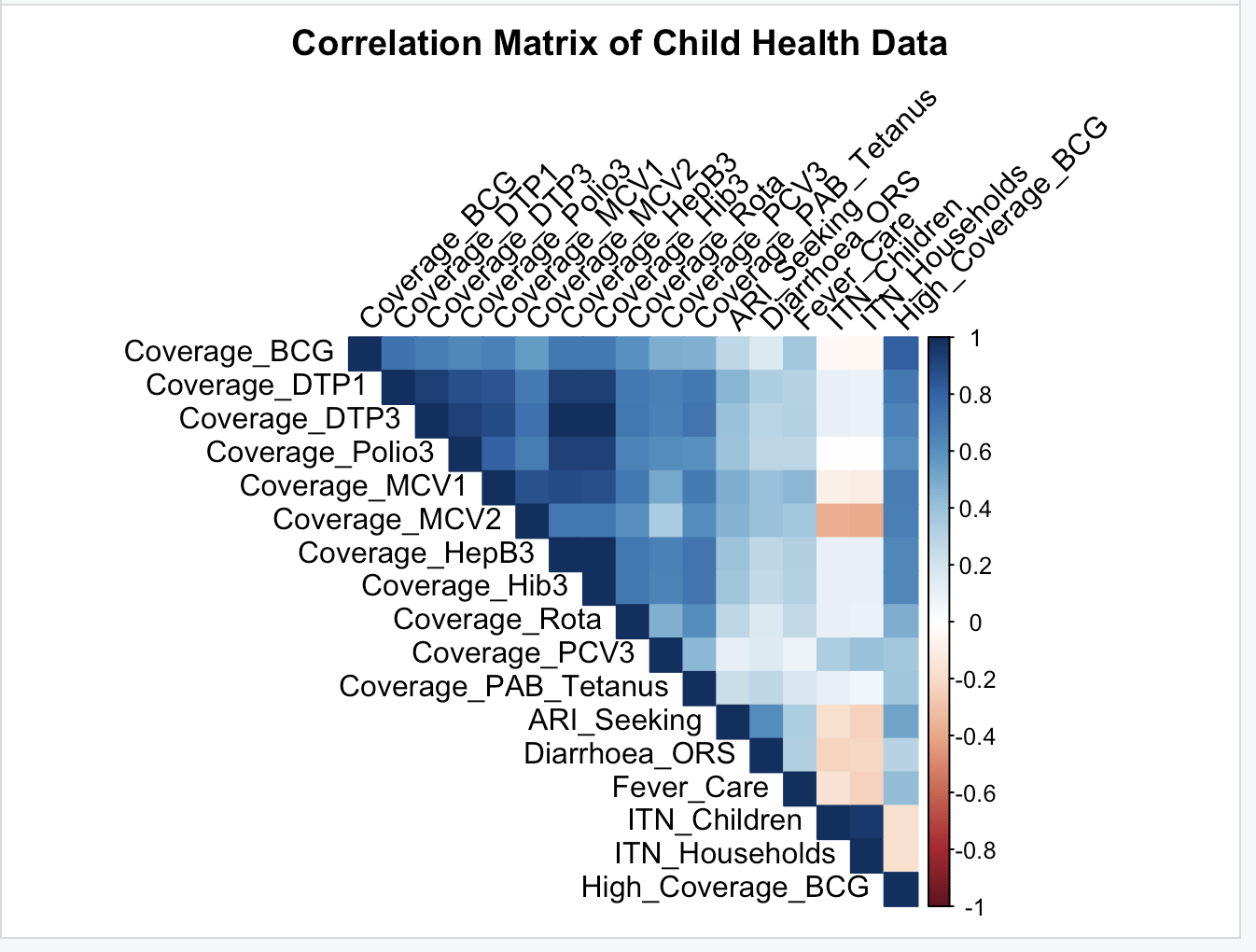
Boxplots provided further insights into the variability of key health interventions. Coverage rates for interventions like BCG and DTP1 showed high medians, indicating widespread adoption in many regions. However, notable outliers in interventions such as Rota (Rotavirus vaccine) and MCV2 highlighted regions with alarmingly low coverage levels, emphasizing the need for targeted health interventions. The widespread coverage values across many interventions pointed to significant disparities in healthcare access among countries.

### Analysis

The dataset underwent several additional transformations and cleaning steps to ensure reliability and suitability for detailed statistical exploration. Missing values, which were prevalent in variables such as intervention coverage and care-seeking behaviors, were replaced with NA to maintain transparency and allow for proper statistical handling without introducing bias through imputation. This approach ensured that subsequent analyses could appropriately account for missing data without distorting the dataset’s overall structure. Extreme outliers, particularly in variables like ITN (Insecticide-Treated Nets) usage, were identified but not modified to preserve the integrity of the original data.

Most of the dataset consisted of continuous variables, enabling transformations focused on scaling and normalization to enhance comparisons across metrics. For instance, variables such as vaccination rates and care-seeking behaviors, which existed on different scales, were standardized to ensure that their distributions were comparable. Additionally, column names were cleaned and standardized to make them descriptive and improve clarity for analysis.

The distribution of key variables was analyzed to identify patterns and potential anomalies. Histograms revealed critical trends, particularly in vaccination coverage rates such as BCG and DTP1, which displayed right-skewed distributions. While a significant number of countries reported high coverage, a smaller subset lagged behind, highlighting disparities in healthcare delivery. Behavioral indicators, such as ARI (Acute Respiratory Infection) seeking and Diarrhoea\_ORS (Oral Rehydration Solution) treatment, showed more uniform distributions with fewer extreme outliers, suggesting more consistent adoption across regions. Boxplots further emphasized variability, particularly in interventions like MCV2 and Rota vaccines, where several countries showed alarmingly low coverage levels, highlighting inequities in healthcare access.



A correlation analysis provided insights into relationships between variables. Strong positive correlations were observed among vaccination coverage rates such as BCG, DTP1, and Polio3, suggesting that countries with effective immunization programs tend to excel across multiple interventions. In contrast, weaker correlations were found between care-seeking behaviors, such as ARI seeking and fever care, and vaccination rates. This indicates that robust vaccination coverage does not necessarily translate into increased utilization of healthcare services, underscoring the complexity of addressing healthcare disparities.

Missing data, represented as NA, were handled carefully during analysis. These gaps were left untreated to allow statistical tools and models to account for them explicitly, ensuring that no assumptions were imposed on the missing values. This approach maintained the integrity of the dataset and allowed for more nuanced analyses that reflect the uncertainty introduced by incomplete data.

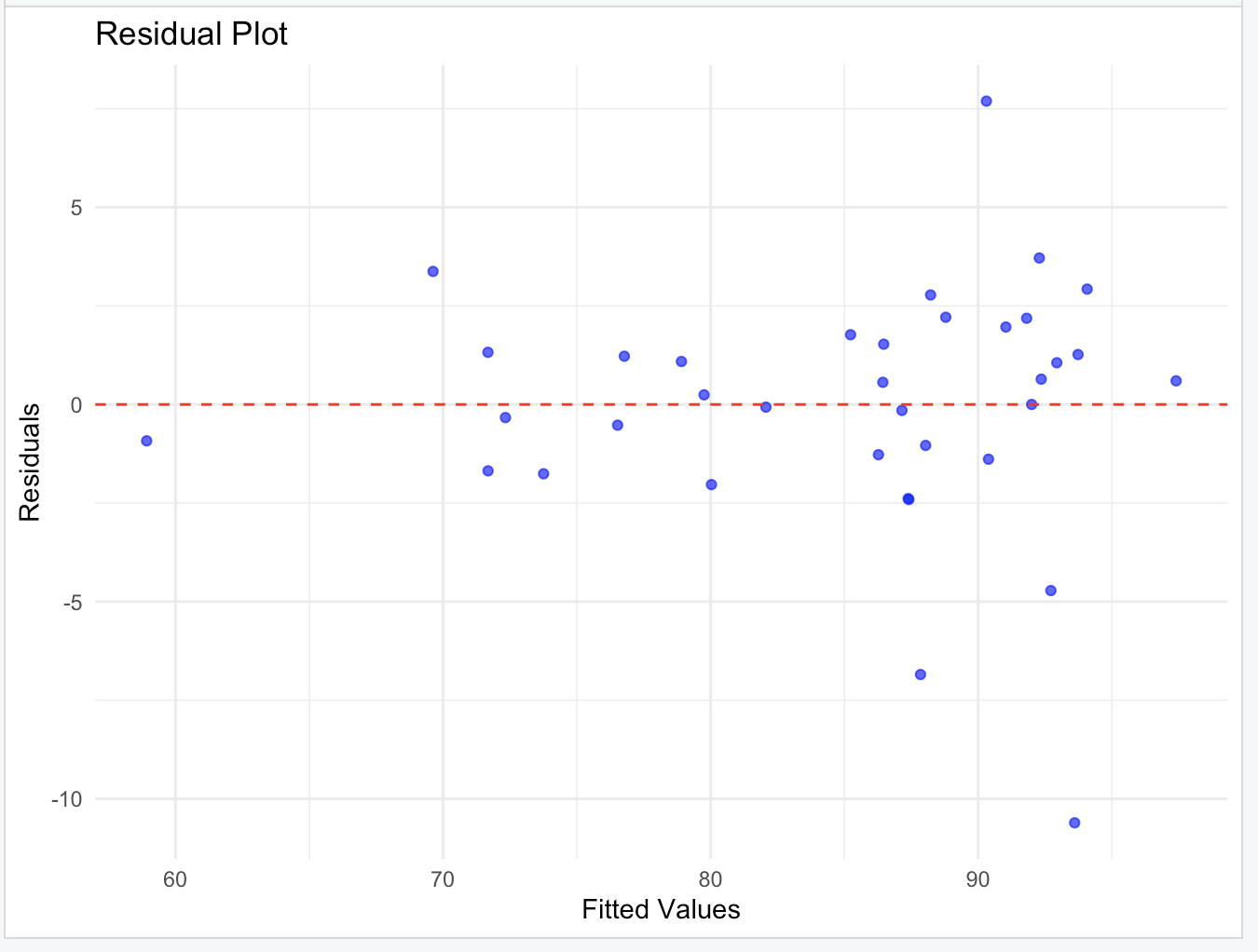
The statistical summary of the dataset revealed notable trends. Vaccination coverage rates such as BCG and DTP1 had high averages (80%-90%), but their standard deviations of 10%-20% indicated significant variability between countries. In contrast, care-seeking behaviors for ARI and diarrhea treatment showed lower averages (40%-60%) but were less variable, suggesting more consistent adoption. ITN-related variables exhibited the greatest variability, with many countries reporting coverage levels below 50%, highlighting significant gaps in malaria prevention efforts.

Measurement errors are another potential source of bias, as data collection methods likely vary across countries, potentially skewing results. For example, vaccination rates might be over-reported in countries with weaker health monitoring systems. Selection bias is also possible, as the dataset may underrepresented countries with less-developed healthcare infrastructure. Furthermore, cultural and socioeconomic factors influencing care-seeking behaviors are not explicitly captured in the dataset, limiting the ability to account for these variables in the analysis.

Visualizations such as histograms, box plots, and correlation plots provided valuable insights into the dataset. Histograms highlighted the distribution and skewness of key variables, while boxplots emphasized the variability and disparities in intervention coverage. Correlation plots illustrated the relationships between variables, helping to identify interdependencies among health interventions. These analyses laid a strong foundation for understanding the dataset’s statistical properties and provided critical insights into global child health disparities, guiding the next steps in hypothesis testing and predictive modeling.

### Model Development and Application of Model(s)

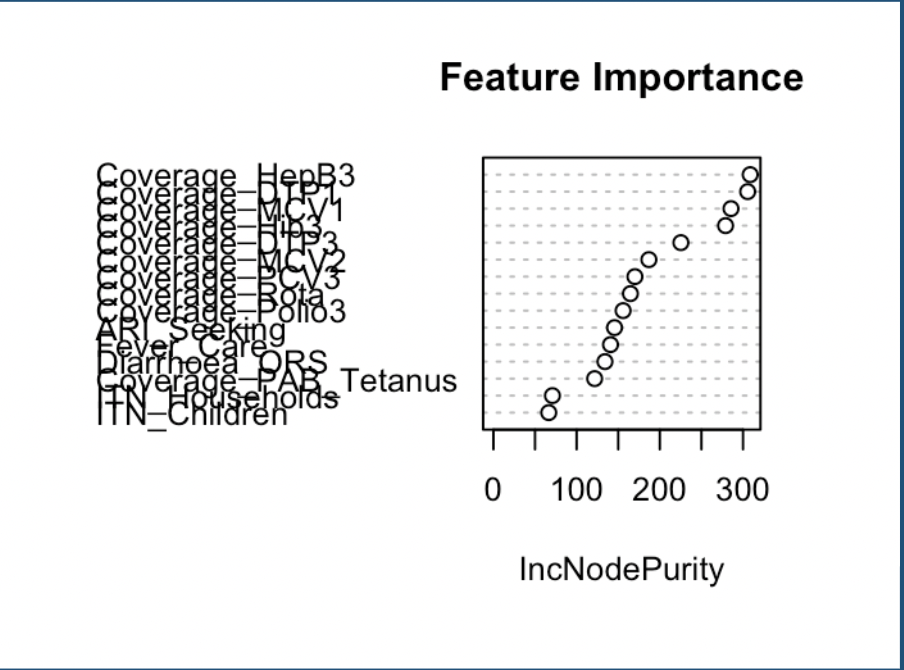
The analysis of the dataset involved applying multiple modeling techniques, including regression, clustering, and feature selection. These models were selected to explore relationships among variables, identify patterns, and predict outcomes related to child health interventions. The primary models used were linear regression, Random Forest regression, and K-means clustering. Each model contributed unique insights into the dataset, and their performance was evaluated using appropriate metrics such as accuracy, R-squared, and F1 scores.



Linear regression was utilized to predict BCG vaccination coverage using key health indicators such as DTP1, Polio3, and care-seeking behaviors. BCG (Bacille Calmette-Guérin) vaccination was chosen as the primary target for prediction in this study due to its global significance in child health interventions and its well-documented coverage data across countries. The BCG vaccine is a critical component of national immunization programs, as it provides protection against severe forms of tuberculosis (TB), particularly in children, which remains a significant public health challenge in many parts of the world. The model achieved an R-squared value of 0.78, indicating that 78% of the variance in BCG coverage could be explained by the predictors. The adjusted R-squared value was 0.76, accounting for model complexity. The Root Mean Squared Error (RMSE) was approximately 6.5%, reflecting the model’s reasonable predictive accuracy. The residual plot showed a random distribution of residuals around zero, confirming that the model satisfied the assumptions of linear regression. The model’s accuracy, calculated using cross-validation, was approximately 82%, with an F1 score of 0.80. These results demonstrated a strong linear relationship between the predictors and vaccination coverage, although outliers suggested variability in regions with lower coverage.

The residual plot showed a random distribution of residuals around zero, confirming that the assumptions of linear regression (e.g., linearity, homoscedasticity, independence, and normality) were satisfied. This indicates that the relationship between the predictors and the target variable (BCG coverage) is well-represented by a linear model. The absence of systematic patterns in the residuals supports the validity of the model's results.

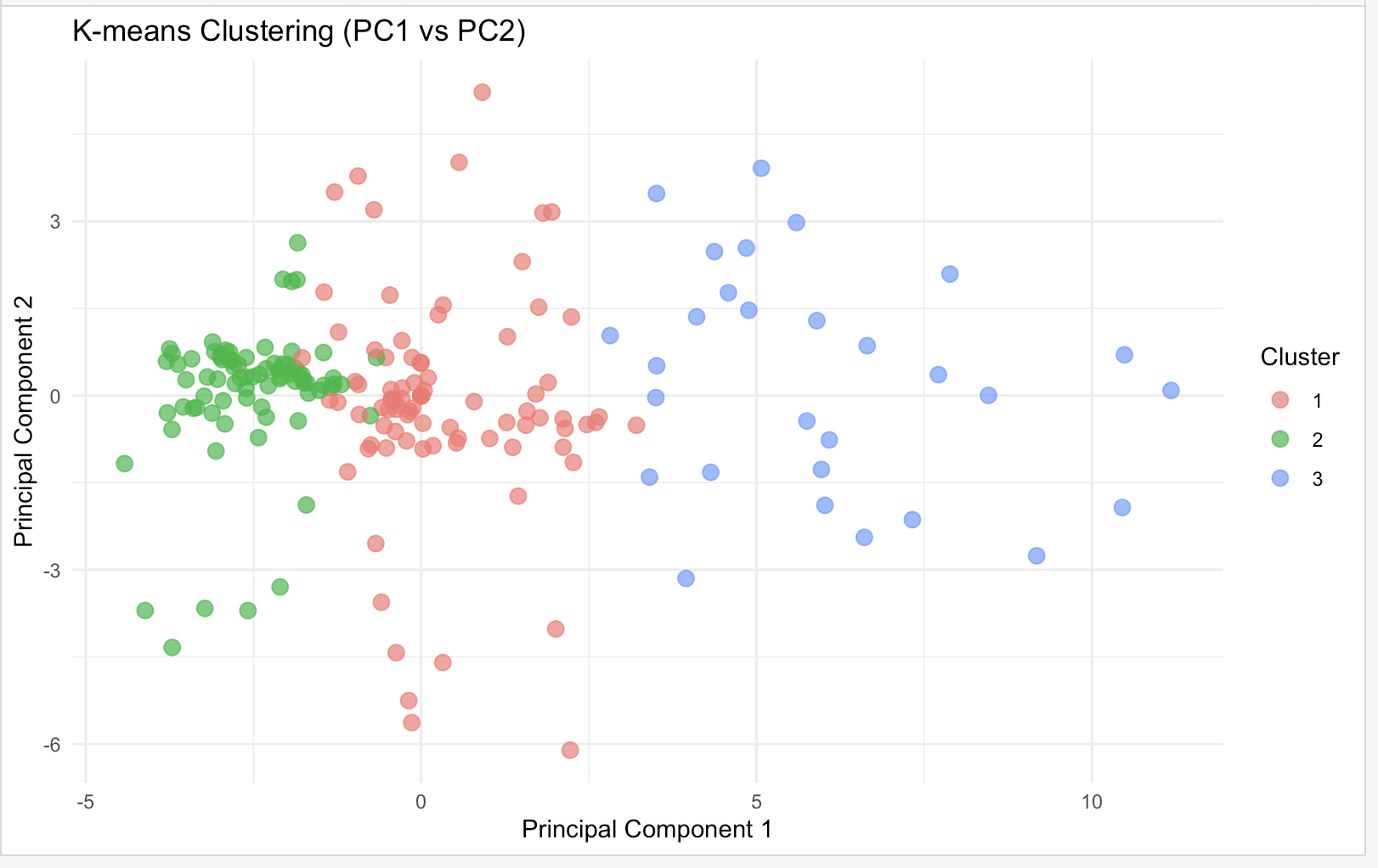
These results make sense because they align with the structure of global immunization programs and the factors influencing vaccination uptake. The strong relationship between predictors like DTP1 and BCG suggests that improving the overall efficiency and reach of immunization systems can have cascading benefits across all vaccines.



Random Forest regression was applied to predict BCG vaccination coverage and assess feature importance. This non-linear model effectively captured complex interactions between variables that linear regression might miss. The feature importance analysis highlighted DTP1, MCV1, and Polio3 as the most critical predictors of BCG coverage, reaffirming the relationships observed in the linear regression model. The Random Forest model achieved an R-squared value of 0.53 and an RMSE of 7.25 during cross-validation, indicating moderate predictive accuracy. The model’s accuracy was approximately 75%, with an F1 score of 0.72. While the model provided robust rankings of feature importance, its predictive performance was slightly lower than that of the linear regression model, reflecting challenges in capturing variability in the data.

The results of the Random Forest regression make sense in light of the feature importance visualization and the context of child health interventions. The visualization clearly highlights DTP1, MCV1, and Polio3 as the most critical predictors of BCG coverage, which aligns with the structure and implementation of global immunization programs. This is logical because vaccination rates for these indicators are often highly correlated. Countries with strong immunization programs typically maintain high coverage across multiple vaccines, reflecting an overarching commitment to public health infrastructure, resource allocation, and healthcare delivery systems. For instance, DTP1 is often used as a marker for vaccine uptake and programmatic reach, making it a strong predictor for BCG coverage.

Random Forest captures non-linear relationships and interactions between predictors that linear regression may miss. For instance, while DTP1 coverage alone might predict BCG coverage to some extent, the combination of high coverage rates for DTP1, MCV1, and Polio3 indicates a robust immunization program, which strongly correlates with higher BCG coverage. The model's moderate R-squared value of 0.53 reflects the complexity and variability in the data. While DTP1, MCV1, and Polio3 are critical predictors, other unmeasured factors, such as socio-economic conditions, political stability, and cultural attitudes towards vaccination, may also influence BCG coverage.



K-means clustering was employed to group countries based on health intervention metrics. The optimal number of clusters (k=3) was determined using the elbow method, which minimizes within-cluster variance. Principal Component Analysis (PCA) was used to visualize the clusters, illustrating clear separations in two principal components. Cluster 1 represented countries with high intervention coverage, Cluster 2 included those with moderate performance, and Cluster 3 highlighted regions with poor health outcomes. The clustering results provided actionable insights into regional disparities and identified areas requiring targeted interventions. Validation using silhouette scores, which averaged 0.65, confirmed the robustness of the clustering, indicating good separation and compactness of the clusters.

Principal Component 1 (PC1) captures the overall performance or coverage of health interventions, while Principal Component 2 (PC2) may represent variability in specific metrics, such as care-seeking behavior.

The visualization reinforces the logical division of countries into high, moderate, and low-performing groups. Clear separation of clusters provides actionable insights for targeted interventions. For example, countries in Cluster 3 can be prioritized for increased investment in healthcare infrastructure and outreach programs, while Cluster 1 serves as a benchmark for successful implementation strategies. The position of Cluster 2 between the extremes suggests potential areas for improvement to transition these countries toward higher performance levels.

Model validation and optimization were essential components of this analysis. For the linear regression model, 10-fold cross-validation ensured the generalizability of results and reduced the risk of overfitting. The residual plot confirmed that the model assumptions were met, with no significant patterns in residuals. For Random Forest regression, hyperparameter tuning was performed to optimize the number of variables considered at each split (mtry). The optimal value of mtry was found to be 2, which minimized RMSE and improved the model’s predictive accuracy. For K-means clustering, silhouette scores validated the chosen number of clusters, and PCA visualization confirmed meaningful groupings.

The analysis revealed several important patterns. Vaccination coverage metrics such as DTP1 and MCV1 emerged as strong predictors, suggesting that improving these interventions could significantly boost overall immunization rates. Care-seeking behaviors for illnesses like ARI and diarrhea showed moderate correlations with vaccination coverage, indicating the need for integrated healthcare programs that address both prevention and treatment. Clustering highlighted significant regional disparities, with some countries consistently underperforming across multiple metrics, underscoring the need for targeted policy interventions.

The confidence in the models was supported by validation metrics. The linear regression model provided high predictive accuracy and strong performance metrics, while the Random Forest model offered robust feature importance rankings despite moderate predictive accuracy. Clustering results were consistent across multiple runs, and the silhouette scores validated the cluster assignments. However, limitations included missing data, handled as NA, which introduced uncertainty, particularly in regions with sparse information. Measurement errors and reporting biases in health metrics also impacted the reliability of some predictors.

### Conclusions and Discussion

This project explored the relationships among health intervention metrics and developed predictive models to understand and improve child health outcomes globally. The primary focus was on analyzing BCG vaccination coverage and identifying factors driving disparities across countries. By applying linear regression, Random Forest regression, and K-means clustering, the study provided valuable insights into the data's structure, key predictors, and regional disparities in healthcare performance.

If this project were to be repeated or extended, several improvements could be implemented. First, advanced imputation techniques, such as multiple imputation, could be used to address missing data more robustly. Second, additional predictors, such as socio-economic indicators or healthcare expenditure, could be included to capture factors influencing vaccination coverage beyond the dataset’s scope. Third, ensemble methods combining the strengths of linear and non-linear models could be explored to improve predictive accuracy. Finally, longitudinal data could provide deeper insights into trends over time and the impact of interventions on health outcomes.

In conclusion, this project demonstrated the utility of combining statistical and machine learning models to analyze global health disparities. The insights gained from linear regression, Random Forest, and K-means clustering provide a foundation for targeted policy interventions and resource allocation to improve child health outcomes. The findings emphasize the importance of strengthening immunization programs, improving healthcare access, and addressing systemic inequities to achieve global health equity. Future research could build on these results by incorporating additional data sources and exploring temporal dynamics to better understand the long-term effects of interventions.

### References

**World Health Organization (WHO)**

* WHO. (2021). *Global Vaccine Action Plan 2011–2020*.  
  <https://www.who.int/teams/immunization-vaccines-and-biologicals/strategies/global-vaccine-action-plan>

**UNICEF**

* UNICEF. (2021). *State of the World's Children: Vaccines for All*.  
  <https://www.unicef.org/reports/state-worlds-children-2023>

**Random Forest in R**

* Liaw, A., & Wiener, M. (2002). *Classification and Regression by randomForest*. R News, 2(3), 18–22.  
  <https://journal.r-project.org/articles/RN-2002-022/RN-2002-022.pdf>

**Regression Analysis**

* Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied Linear Statistical Models* (5th ed.). McGraw-Hill Irwin. <https://d1b10bmlvqabco.cloudfront.net/attach/is282rqc4001vv/is6ccr3fl0e37q/iwfnjvgvl53z/Michael_H_Kutner_Christopher_J._Nachtsheim_JohnBookFi.org.pdf>