Reall Household Income and Affordability Calculator

Methodology Paper: Kenya 2016

This paper lays out the methodology used to develop the 2016 Kenya data for Reall's Household Income and Affordability Calculator, which can be found at www.reall.net/calculator.

The Integrated Household Budget Survey (IHBS) is produced by the Kenya National Bureau of Statistics. IHBS is a nationally-representative household survey focused on presenting income and consumption expenditure data, with the current round covering 21,774 households. The full data and descriptions are freely available from the Kenya National Data Archive[1], with further information also available from the International Household Survey Network. [2]

The IHBS focuses on consumption expenditure, so this is used as a proxy for income. This is generally seen as good practice, providing more reliable responses from participants. However, it may also result in slightly lower figures than would have been collected for income, particularly for wealthier households.

Data is split into separate data files for different sections of the survey. Key sections of the IHBS survey used for this work are Consumption_aggregate, which provides all household expenditure, household sizes and weightings; and HH_Members_Information, which provides employment data for each household member. Names and coding for Kenya's counties can be found on the Data Description tab under 'HH_Information', [county]. [3]

Methodology

The following actions were undertaken on these datasets:

Creating Dataset

- HH_Members_Information
    - Creating new [hhcode] (household code) by pasting together [clid] ('Cluster ID') and [hhid] ('Household ID')
    - Adjust [d04_1], [d04_2], and [d04_3] to give each a value of 1 if they currently contain 'A' to 'F', and 0 otherwise
    - Create new [d04] by summing [d04_1], [d04_2] and [d04_3]
    - Create new [is_working] column with a value of 1 if any of [d02_1], [d02_2], [d02_3], [d02_4], [d02_5] or [d02_6] have a value of 1, and 0 otherwise
    - Create new [is_earner] column with a value of 1 if [is_working] plus [d04] is greater than 0, and 0 otherwise
    - Group all data by [HHID] (household ID), summing number [is_earner] household members in a new [no_of_earners] column
- Consumption_aggregate

[1] https://statistics.knbs.or.ke/nada/index.php/catalog/13/study-description
[2] https://catalog.ihsn.org/catalog/7432/study-description
[3] https://catalog.ihsn.org/catalog/7432/study-description

- Creating new [hh_exp] (household expenditure) column, by multiplying [padqexp] ('Monthly per adult equivalent total consumption expenditure (deflated)') by [ctry_adq] ('Sum total of adult equivalent scales (country-specific)'
- Converting [resid] to new [urban_rural] column, where 1 is 'rural' and 2 is 'urban'
- Creating new [hhcode] (household code) by pasting together [clid] ('Cluster ID') and [hhid] ('Household ID')
- Merger of 'county_value' by [county]
- Merger of 'HH_Members_Information' by [hhcode]

```
hh_dataset <- hh_dataset %>%
        mutate(
                hhcode = paste (clid,hhid)
                ) %>%
        mutate(
                d04_1 = ifelse(d04_1 == 'A', 1, ifelse(d04_1 == 'B', 1, ifelse(d04_1 == 'C', 1,
                ifelse(d04_1 == 'D', 1, ifelse(d04_1 == 'E', 1, ifelse(d04_1 == 'F', 1, 0))))),
                d04_2 = ifelse(d04_2 == 'A', 1, ifelse(d04_2 == 'B', 1, ifelse(d04_2 == 'C', 1,
                ifelse(d04_2 == 'D', 1, ifelse(d04_2 == 'E', 1, ifelse(d04_2 == 'F', 1, 0))))),
                d04_3 = ifelse(d04_3 == 'A', 1, ifelse(d04_3 == 'B', 1, ifelse(d04_3 == 'C', 1,
                ifelse(d04_3 == 'D', 1, ifelse(d04_3 == 'E', 1, ifelse(d04_3 == 'F', 1, 0))))))
                ) %>%
        mutate(
                is_working = ifelse(d02_1 == 1,1, ifelse(d02_2 == 1,1, ifelse(d02_3 == 1,1,
                ifelse(d02_4 == 1,1,ifelse(d02_5 == 1 ,1, ifelse(d02_6 == 1, 1, 0)))))) %>%
        mutate (d04 = d04_1 + d04_2 + d04_3) %>%
        mutate (is_earner = ifelse(d04 + is_working > 0,1,0)) %>%
        group_by(hhcode) %>%
        summarise(no_of_earners = sum(is_earner)) %>%
        ungroup()

consumption_dataset <- consumption_dataset %>%
        mutate(
                HH_exp = padqexp * ctry_adq,
                Percent_Income_Spent_on_Housing = (padqrent / HH_exp) * 100,
                urban_rural = if_else(resid == 1, "Rural", "Urban"),
                hhcode = paste (clid,hhid)
                ) %>%
        left_join(counties_code, by = "county") %>%
        left_join(hh_dataset, by = c("hhcode")) %>%
        distinct()
```

Calculating Sample Sizes

- Grouping dataset by [county_value] to create sample sizes for each county
- Grouping dataset by [county_value] and [urban_rural] to create separate urban and rural sample sizes for each county

```
sample_size_kenya <- consumption_dataset %>%
```

```
        group_by(county_value, urban_rural, sample_year = 2016) %>%
        summarise(sample_size = n())

sample_size_kenya_u_r <- consumption_dataset %>%
        group_by(county_value, urban_rural ="All", sample_year = 2016) %>%
        summarise(sample_size = n())
```

Calculating Percentiles

- Multiplying each [hhcode] record by [weight] to create full weighted dataset
- Grouping data by [county_value], and extracting data for records at 1% increments, creating figures for each percentile of every county
- Repeating the step above but also grouping by [urban_rural], creating separate urban and rural percentile figures for each county

```
# Function to calculate quantiles and related statistics
calculate_quantiles <- function(data, quantiles) {
        do.call(rbind, lapply(quantiles, function(q) {
                data.frame(
                        Quantile = q * 100,  # Convert quantile probability to percentage
                        HH_exp = quantile(data$HH_exp, probs = q, na.rm = TRUE)
                        )
                }))
        }

all_data_kenya <- consumption_dataset %>%
        uncount(weights = as.integer(weight / 100)) %>%
        rename(HH_size = hhsize)

# Calculate quantiles for each location
kenya_location_quantiles <- all_data_kenya %>%
        group_by(Country = "Kenya", urban_rural, City = county_value, Year = 2016) %>%
        group_modify(~ calculate_quantiles(.x, quantile_probs)) %>%
        ungroup()

# Additional mixed urban/rural quantiles for each location
quantile_probs <- seq(0.01, 0.99, by = 0.01)
kenya_location_quantiles_u_r <- all_data_kenya %>%
        group_by(Country = "Kenya", urban_rural = "All", City = county_value, Year = 2016)
        %>%
        group_modify(~ calculate_quantiles(.x, quantile_probs)) %>%
        ungroup()

kenya_location_quantiles["Location"] <- "All"
kenya_location_quantiles_u_r ["Location"] <- "All"
```

Creating Final Dataset

- Define and align common columns to enable merging of quantiles datasets
- Combine relevant 'quantiles' and 'quantiles_u_r' datasets into a single 'summary_dataset'
- Define and align common columns to enable merging of sample size datasets
- Combine relevant 'sample_size' datasets into a single 'combined_sample_size'
- Join 'combined_sample_size' dataset to 'summary_dataset'
- Create a 'national_aggregated' version of 'summary_dataset' by grouping all household data by country
- Join 'national_aggregated' dataset to 'summary_dataset'

```
# Define common columns for final summary
common_columns <- c("Country", "urban_rural", "Location", "City", "Year", "Quantile",
"HH_exp", "no_of_earners", "HH_size", "Percent_Income_Spent_on_Housing")

# Function to align columns across datasets
align_columns <- function(df, common_cols) {
        df %>%
        mutate(across(setdiff(common_cols, colnames(df)), ~ NA)) %>%
        select(all_of(common_cols))
        }

# Align columns and combine all datasets
Kenya_location_quantiles <- align_columns(Kenya_location_quantiles, common_columns)
Kenya_location_quantiles_u_r <- align_columns(Kenya_location_quantiles_u_r,
common_columns)

# Combine all country datasets into a single summary dataset
summary_dataset <- bind_rows(Kenya_location_quantiles, Kenya_location_quantiles_u_r)

sample_size_Kenya <- sample_size_Kenya %>%
        rename(Location = state, sample_size = sample_size) %>%
        mutate(Country = "Kenya")

sample_size_Kenya_u_r <- sample_size_Kenya_u_r %>%
        rename(Location = state, sample_size = sample_size) %>%
        mutate(Country = "Kenya")

# Combine the sample size tables into one
combined_sample_size <- bind_rows(
        sample_size_Kenya,
        sample_size_Kenya_u_r,

summary_dataset <- summary_dataset %>%
        left_join(combined_sample_size, by = c ("Country", "urban_rural", "Location", "City",
        "Year"))

# Aggregate data for the national level by combining all states within each country
national_aggregated <- summary_dataset %>%
        group_by(Country, Year, Quantile,urban_rural) %>%
```

```
        summarise(
                sample_size = sum(sample_size),
                HH_exp = mean(HH_exp, na.rm = TRUE),
                no_of_earners = mean(no_of_earners, na.rm = TRUE),
                HH_size = mean(HH_size, na.rm = TRUE)
                ) %>%
        ungroup() %>%
        mutate(Location = "All", City = "All")

# Combine both the state-level and national-level data
final_dataset <- bind_rows(state_aggregated, national_aggregated,summary_dataset)
```

Calculating Inflation

All data is inflated using median annual inflation rates from 2010-23. Median rates were used rather than actual figures to help compensate for large-scale inflation across many economies in 2022 and 2023.

Consumer Price Index inflation rates were taken from the World Bank[4] and consisted of the following figures, creating a final median rate of 6.44%.

CPI	Kenya
2000	9.980025
2001	5.738598
2002	1.961308
2003	9.815691
2004	11.62404
2005	10.31278
2006	14.45373
2007	9.75888
2008	26.23982
2009	9.234126
2010	3.961389
2011	14.02249
2012	9.377767
2013	5.717494
2014	6.878155
2015	6.582174
2016	6.297158
2017	8.005723
2018	4.68982
2019	5.23586
2020	5.404815
2021	6.110909

[4] https://data.worldbank.org/indicator/FP.CPI.TOTL.ZG

2022	7.659863
2023	7.671396