A HOME FOR HUMAN POTENTIAL

Reall Household Income and Affordability Calculator

Methodology Paper: Kenya 2021

This document outlines the methodology used to process the 2021 Kenya Household Integrated Budget Survey (KHIBS) data for inclusion in Reall's Household Income and Affordability Calculator. The calculator is available at www.reall.net/calculator.

Background and Data Source

The **Kenya Continuous Household Survey (KCHS) 2021**[1], conducted by the **Kenya National Bureau of Statistics (KNBS)** [2]in collaboration with the State Department for Planning, is the primary data source for this analysis. The survey was funded by the Government of Kenya and supported by The World Bank. It forms part of Kenya's official programme of integrated household surveys, designed to generate regular, nationally representative statistics on welfare, poverty, and household living standards.

The survey adopts a **stratified, multi-stage cluster sampling design** based on the 2019 Kenya Population and Housing Census. This ensures comprehensive coverage at the **national, county, urban, and rural levels**, capturing the diversity of Kenya's socio-economic landscape. Data were collected through **face-to-face interviews** conducted by trained KNBS enumerators.

The KCHS provides detailed household- and individual-level data, covering consumption aggregates, expenditure patterns, housing costs, and demographic characteristics. Key documentation for the survey includes the survey questionnaire, a poverty report, and a technical manual, all accessible via the KNBS National Data Archive (KeNADA).

For this methodology, two household-level microdata files were used:

- **consagg_microdata.dta –** containing household-level consumption aggregates, including expenditure and rent.
- **households_microdata.dta –** containing household demographics, household size, and survey weights.

Additionally, a county-level mapping file (**kenya_counties.csv**) was used to attach county names to household records. A unique household identifier (hhcode) was created by concatenating the cluster ID (clid) and household ID (hhid) to ensure consistency across datasets.

[1] https://statistics.knbs.or.ke/nada/index.php/catalog/123/related-materials

[2] https://statistics.knbs.or.ke/nada/index.php/home

Dataset Creation

Input files

- consagg_microdata.dta – household-level consumption aggregates
- households_microdata.dta – household-level demographics and weights
- kenya_counties.csv – county code-to-name mapping

Variables extracted and transformations performed

- **HH_exp**: Adjusted household expenditure (padqexp * adq_scale)
- **Percent_Income_Spent_on_Housing**: (padqrent / HH_exp) * 100
- **HH_size**: Household size (hhsize)
- **no_of_earners**: Set to NA (not available in KHIBS 2021 aggregate file)
- **urban_rural**: Based on resid (1 = Rural, 2 = Urban)
- **weight_hh**: Household survey weight
- **hhcode**: Concatenation of clid and hhid for unique household ID
- **City**: County name from kenya_counties.csv
- **Location**: Assigned as "All"
- **Country**: Kenya
- **Year**: 2021

```
#Kenya 2021 Data Processing

# === Step 0: Load datasets ===


consump_kenya_2021 <-
read_dta("C:/Users/Pratyusha/Documents/Reall/Data/Datasets/Kenya/2021/consagg_microdata.dta")

hh_kenya_2021 <-
read_dta("C:/Users/Pratyusha/Documents/Reall/Data/Datasets/Kenya/2021/households_microdata.dta")

county_kenya2021_map <-
read_csv("C:/Users/Pratyusha/Documents/Reall/Data/Datasets/Kenya/kenya_counties.csv")


# === Step 1: Merge county names into household data ===

hh_kenya_2021 <- hh_kenya_2021 %>%

  left_join(county_kenya2021_map, by = c("county" = "county")) %>%

  mutate(hhcode = paste0(clid, hhid))


# === Step 2: Process consumption and join ===

kenya2021 <- consump_kenya_2021 %>%

  mutate(

    hhcode = paste0(clid, hhid),

    HH_exp = padqexp * adq_scale,

    Percent_Income_Spent_on_Housing = (padqrent / HH_exp) * 100,

    no_of_earners = NA_real_,

    urban_rural = if_else(resid == 1, "Rural", "Urban")

  ) %>%

  left_join(hh_kenya_2021 %>% select(hhcode, county_value), by = "hhcode") %>%

  rename(HH_size = hhsize, weight_hh = weight_hh, City = county_value) %>%

  mutate(Country = "Kenya", Year = 2021, Location = "All") %>%

  distinct()
```

Expanding Dataset by Weights

To reflect actual population distributions, the dataset was expanded by household weights (weight_hh). Each household record was replicated according to its normalized weight (scaled by 1/100). This ensured that quantile calculations represented Kenya's weighted population, not just the survey sample.

```
# === Step 4: Expand by weights ===
expanded_kenya2021 <- kenya2021 %>%
  uncount(weights = as.integer(weight_hh / 100))
```

Calculating Sample Sizes

Two sets of sample sizes were calculated:

1. **Urban/Rural disaggregated** – counts of households by county (City) and urban–rural classification.
2. **All households combined** – counts of households by county, aggregating across urban and rural.

These were later merged into the quantile summary dataset.

```
# === Step 3: Sample sizes ===
sample_size_kenya_2021 <- kenya2021 %>%
  group_by(Country, Year, Location, City, urban_rural) %>%
  summarise(sample_size = n(), .groups = "drop")


sample_size_all_kenya_2021 <- kenya2021 %>%
  group_by(Country, Year, Location, City, urban_rural = "All") %>%
  summarise(sample_size = n(), .groups = "drop")
```

Quantile Computation

Percentile values (1st to 99th) were computed for each county and urban/rural grouping using the expanded dataset. For each quantile q, the following statistics were derived:

- **HH_exp** (Household expenditure)

- **HH_size** (Household size)
- **Percent_Income_Spent_on_Housing**

Quantiles were calculated separately for:

- **Urban/Rural disaggregated** (kenya_quantiles_2021)
- **All households combined** (kenya_quantiles_2021_all)

```
# === Step 6: Compute quantiles ===
kenya_quantiles_2021 <- expanded_kenya2021 %>%
  group_by(Country, Location, City, urban_rural, Year) %>%
  group_modify(~ calculate_quantiles(.x, quantile_probs)) %>%
  ungroup()

kenya_quantiles_2021_all <- expanded_kenya2021 %>%
  group_by(Country, Location, City, urban_rural = "All", Year) %>%
  group_modify(~ calculate_quantiles(.x, quantile_probs)) %>%
  ungroup()
```

Final Dataset Preparation

Quantile and sample size datasets were merged to ensure that each quantile summary included the correct sample size information. Specifically:

1. **Merge quantile datasets with sample sizes**:
    - The kenya_quantiles_2021 dataset was merged with sample_size_kenya_2021.
    - The kenya_quantiles_2021_all dataset (urban + rural combined) was merged with sample_size_all_kenya_2021.
2. **Combine merged datasets**:
   The two merged datasets were appended together into a single file (kenya_final_2021).
3. **Select and align columns**:
   Variable names and column order were standardized using the align_columns() function to ensure compatibility with other country datasets.
4. **Append to global summary dataset**:
   The cleaned and aligned Kenya 2021 file was added into the cross-country summary_dataset

```
# === Step 7: Combine and export ===

kenya_final_2021 <- bind_rows(kenya_quantiles_2021, kenya_quantiles_2021_all) %>%

  left_join(bind_rows(sample_size_kenya_2021, sample_size_all_kenya_2021), by = c("Country",
"Year", "Location", "City", "urban_rural")) %>%

  select(Country, Location, Year, Quantile, urban_rural, sample_size,

         HH_exp, no_of_earners, HH_size, Percent_Income_Spent_on_Housing, City)

# write_csv(kenya_final, "kenya_quantile_summary_2021.csv")

# === Step 8: Integrate into Combined Dataset ===

kenya_final_combined_aligned_2021 <- align_columns(kenya_final_2021, common_columns)

summary_dataset <- bind_rows(summary_dataset, kenya_final_combined_aligned_2021)

# combined_sample_size <- bind_rows(

#   combined_sample_size,

#   sample_size_kenya_2021,

#   sample_size_all_kenya_2021 )
```

```
combined_sample_size <- bind_rows(

  sample_size_india,  sample_size_india_u_r,  sample_size_uganda,  sample_size_uganda_u_r,

  sample_size_nigeria,  sample_size_nigeria_u_r,   sample_size_pakistan,
sample_size_pakistan_u_r,  sample_size_kenya,   sample_size_kenya_u_r,

  sample_size_kenya_2020,  sample_size_kenya_2020_u_r,  sample_size_india_2022,

  sample_size_india_u_r_2022,  sample_size_ph_2023,  sample_size_ph_u_r_2023,
sample_size_ph_2018,  sample_size_ph_u_r_2018,  sample_size_rw,  sample_size_rw_u_r,
sample_size_mw, sample_size_mw_u_r, sample_size_ci, sample_size_ci_u_r,sample_size_bf,

 sample_size_bf_u_r, sample_size_kenya_2021,  sample_size_all_kenya_2021

) summary_dataset <- summary_dataset %>%

  left_join(combined_sample_size, by = c("Country","urban_rural","Location","City","Year"))
```

Aggregation and Exclusions

- **County-level aggregation**: Kenya's data were retained at county and quantile levels.
- **National-level aggregation**: All counties collapsed into a national summary.

Unlike other countries, Kenya was excluded from the generic state_aggregated step because its processing was already handled separately, ensuring no duplication of results.

The final output files were:

- Summary_Dataset_v9.csv – main dataset with all countries including Kenya 2021
- sample_size_dataset_v9.csv – harmonized sample sizes across all countries

```
state_aggregated <- summary_dataset %>%
  group_by(Country, Location, 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),
    Percent_Income_Spent_on_Housing = mean(Percent_Income_Spent_on_Housing, na.rm =
TRUE)
  ) %>%
  filter (Country != "Kenya") %>%
  ungroup()%>%
  mutate(City = "All")
```

```
# National-Level Aggregation

# 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)

    #Percent_Income_Spent_on_Housing = mean(Percent_Income_Spent_on_Housing, 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)

# Write the final dataset to a CSV file

write.csv(final_dataset, file = "C:/Users/Pratyusha/Documents/Reall/Summary_Dataset_v9.csv", 
row.names = FALSE)

write.csv(combined_sample_size, file = 
"C:/Users/Pratyusha/Documents/Reall/sample_size_dataset_v9.csv", row.names = FALSE)
```

Calculating Income Growth

All household income data for the Kenya was projected using GDP per capita growth rates.

For the historical period 2016–2024, actual annual GDP per capita growth rates (sourced from the World Bank) [3] were applied directly to the dataset. For the forward-looking period 2025–2040, a median annual growth rate of 2.269% was used.

[3] https://data.worldbank.org/indicator/NY.GDP.PCAP.KD.ZG

Year	GDP per capita growth rate (%)
2000	-2.38611
2001	0.575577
2002	-2.56736
2003	-0.1948
2004	1.891135
2005	2.702441
2006	3.277155
2007	3.669787
2008	-2.75886
2009	0.232598
2010	4.968113
2011	2.269522
2012	1.876708
2013	1.263311
2014	2.592184
2015	2.655894
2016	1.937948
2017	1.604752
2018	3.524048
2019	3.070048
2020	-2.21032
2021	5.565137
2022	2.862813
2023	3.483358
2024	2.470855
2025	2.269522