Reall Household Income and Affordability Calculator

Methodology Paper: Rwanda 2016

This document outlines the methodology used to process the 2016 Fifth Integrated Household Living Conditions Survey (EICV5) data for use in Reall's Household Income and Affordability Calculator. The calculator is available at www.reall.net/calculator.

Background and Data Source — Rwanda (EICV5 2016/17)

Our analysis draws on the Fifth Integrated Household Living Conditions Survey (EICV5) [1]conducted by Rwanda's National Institute of Statistics of Rwanda (NISR)[2]. EICV5 is a nationally representative, multi-topic household survey designed to produce estimates for all 30 districts and by urban/rural residence. 14,580 households were interviewed in the EICV5 cross-section. In addition, 2,427 panel households were interviewed with the same questionnaire. Fieldwork ran for 12 months (Oct 2016–Oct 2017) and was organized into 10 data-collection cycles to capture seasonality in income and consumption. The design is stratified multi-stage: enumeration areas are selected with probability proportional to size (PPS), followed by household sampling within clusters. The EICV5 program covered a main cross-section, plus panel and VUP panel components; in this study we use the cross-sectional "cs_*" files listed in the official data dictionary. Key modules include detailed expenditure (with annual, 4-week, and 7-day recall), labor/employment, housing, and demographics; survey weights must be applied for representativeness.

[1] https://www.quality.dashboard.statistics.gov.rw/index.php/catalog/82/study-description
[2] https://statistics.gov.rw/

Dataset Creation

Input files used

- cs_S0_S5_Household.dta — household identifiers, region/province/district codes, urban/rural (ur), survey weight (weight).
- cs_S6B_Employement_6C_Salaried_S6D_Business.dta — individual labor module to derive earners.
- cs_S8A1_expenditure.dta, cs_S8A2_expenditure.dta, cs_S8A3_expenditure.dta, cs_S8B_expenditure.dta — expenditure modules with different recall periods.
- Rwanda_region_mapping.xlsx, Rwanda_province_mapping.xlsx, Rwanda_district_mapping.xlsx — lookups to attach **RegionName**, **ProvinceName**, **DistrictName**.

Variables extracted and transformations performed

- **HH_exp:** Monthly household expenditure built from four modules:
- **no_of_earners:** Per household, count of members with s6bq6 == 1 (earner flag) and sum by hhid.
- **urban_rural:** "Urban" if ur == 1, else "Rural".
- **norm_weight:** round(weight / 100).
- **Location:** ProvinceName (from lookup).
- **City:** DistrictName (from lookup).
- **Percent_Income_Spent_on_Housing:** NA (not constructed for this wave).
- Records with missing **HH_exp** or **norm_weight** were dropped.

```
# === Rwanda 2016 Data Integration ===

# Load household-level data

hh <- read_dta("C:/Users/Pratyusha/Documents/Reall/Data/Rwanda/cs_S0_S5_Household.dta") %>%

 select(hhid, region, province, district, ur, weight) %>%

 mutate(

  region = as.numeric(region),

  province = as.numeric(province),

  district = as.numeric(district),

  urban_rural = ifelse(ur == 1, "Urban", "Rural")

 )
```

```
# Load mapping files

region_map <-
read_excel("C:/Users/Pratyusha/Documents/Reall/Data/Rwanda/Rwanda_region_mapping.xlsx")

province_map <-
read_excel("C:/Users/Pratyusha/Documents/Reall/Data/Rwanda/Rwanda_province_mapping.xlsx")

district_map <-
read_excel("C:/Users/Pratyusha/Documents/Reall/Data/Rwanda/Rwanda_district_mapping.xlsx")

# Join mappings to household data

hh <- hh %>%

  left_join(region_map, by = "region") %>%

  left_join(province_map, by = "province") %>%

  left_join(district_map, by = "district")

# Calculate number of earners per household

persons <-
read_dta("C:/Users/Pratyusha/Documents/Reall/Data/Rwanda/cs_S6B_Employement_6C_Salaried_S6D_Business.dta")

earners <- persons %>%

  mutate(is_earner = ifelse(s6bq6 == 1, 1, 0)) %>%

  group_by(hhid) %>%

  summarise(no_of_earners = sum(is_earner, na.rm = TRUE))

# Load and process household expenditure

exp1 <-
read_dta("C:/Users/Pratyusha/Documents/Reall/Data/Rwanda/cs_S8A1_expenditure.dta") %>%

  select(hhid, amount = s8a1q3) %>%

  group_by(hhid) %>%

  summarise(exp_a1 = sum(amount, na.rm = TRUE) / 12)  # annual → monthly

exp2 <-
read_dta("C:/Users/Pratyusha/Documents/Reall/Data/Rwanda/cs_S8A2_expenditure.dta") %>%

  select(hhid, amount = s8a2q3) %>%

  group_by(hhid) %>%

  summarise(exp_a2 = sum(amount, na.rm = TRUE))  # 4 weeks → monthly
```

```
exp3 <-
read_dta("C:/Users/Pratyusha/Documents/Reall/Data/Rwanda/cs_S8A3_expenditure.dta") %>%
  select(hhid, amount = s8a3q3) %>%
  group_by(hhid) %>%
  summarise(exp_a3 = sum(amount, na.rm = TRUE) * 4.33)  # 7 days → monthly
exp4 <-
read_dta("C:/Users/Pratyusha/Documents/Reall/Data/Rwanda/cs_S8B_expenditure.dta") %>%
  select(hhid, amount = s8bq3) %>%
  group_by(hhid) %>%
  summarise(exp_b = sum(amount, na.rm = TRUE) * 4.33)  # 7 days → monthly
exp_all <- list(exp1, exp2, exp3, exp4) %>%
  reduce(full_join, by = "hhid") %>%
  mutate(HH_exp = rowSums(across(starts_with("exp_")), na.rm = TRUE))
# Merge household + earners + expenditure
merged_data <- hh %>%
  left_join(earners, by = "hhid") %>%
  left_join(exp_all, by = "hhid") %>%
  mutate(
    norm_weight = round(weight / 100),
    Country = "Rwanda",
    Year = 2016,
    City = DistrictName,
    Location = ProvinceName,
    Percent_Income_Spent_on_Housing = NA_real_
  ) %>%
  drop_na(HH_exp, norm_weight)
```

Expanding Dataset by Weight

Each row was replicated by its norm_weight value to simulate the weighted population. To reflect actual population distributions, we **expanded** the dataset based on the survey weights (weight).

Each household record was duplicated based on its weight so that the final dataset better approximates real-world population figures.

```
# Expand based on weight
expanded_data <- merged_data %>%
  uncount(weights = norm_weight)
```

Calculating Sample Sizes

We computed the number of valid survey households (i.e., sample size) for each **region-province-urban/rural** group. Two sets of sample sizes were calculated:

1. **Urban/Rural disaggregated**
2. **Combined Urban & Rural (All)**

```
# Sample size summary
sample_size_rw <- merged_data %>%
  group_by(Location, City, urban_rural, Year) %>%
  summarise(sample_size = n(), .groups = 'drop') %>%
  mutate(Country = "Rwanda")
sample_size_rw_u_r <- merged_data %>%
  group_by(Location, City, urban_rural = "All", Year) %>%
  summarise(sample_size = n(), .groups = 'drop') %>%
  mutate(Country = "Rwanda")
```

Calculating Percentiles

Using a defined quantile probability range (1% to 99%), percentile values were computed for:

- HH_exp (household expenditure)
- HH_size (household size)

These were grouped by:

- Location (region)
- City (province)
- urban_rural (Urban / Rural / All)

Each quantile represents a snapshot of affordability for that income segment in a given location and classification.

```
# Compute quantiles
rw_quantiles <- expanded_data %>%
  group_by(Location, City, urban_rural) %>%
  group_modify(~ calculate_quantiles(.x, quantile_probs)) %>%
  ungroup() %>%
  mutate(Country = "Rwanda", Year = 2016)
rw_quantiles_u_r <- expanded_data %>%
  group_by(Location, City, urban_rural = "All") %>%
  group_modify(~ calculate_quantiles(.x, quantile_probs)) %>%
  ungroup() %>%
  mutate(Country = "Rwanda", Year = 2016)
```

Final Dataset Preparation

- Quantile and sample size datasets were merged: The disaggregated and combined quantile datasets (rw_quantiles, rw_quantiles_u_r) were merged with their corresponding sample size datasets (sample_size_rw, sample_size_rw_u_r) using left_join() on Country, Location, City, urban_rural, and Year. This ensured that each quantile summary contained the correct sample_size information.
- Columns were aligned using a common schema: The resulting dataset (rw_final_combined) was passed through align_columns() to match the master schema (common_columns), maintaining consistency across all country datasets.
- Combine relevant sample size datasets: The 2016 sample size data (both disaggregated and combined) were appended to the unified dataset combined_sample_size using bind_rows(), enabling a consolidated view of sampling strength across countries and years.
- Join combined sample size to summary dataset: Finally, summary_dataset was updated by appending rw_final_combined_aligned using bind_rows(), ensuring that Rwanda 2016 data were incorporated into the master affordability summary.

```
# Merge and finalize
rw_final <- rw_quantiles %>%
  left_join(sample_size_rw, by = c("Country", "Location", "City", "urban_rural", "Year")) %>%
  select(Country, Location, Year, Quantile, urban_rural, sample_size,
         HH_exp, no_of_earners, HH_size, Percent_Income_Spent_on_Housing, City)
rw_final_combined <- bind_rows(
  rw_final,
  rw_quantiles_u_r %>%
    left_join(sample_size_rw_u_r, by = c("Country", "Location", "City", "urban_rural", "Year")) %>%
    select(Country, Location, Year, Quantile, urban_rural, sample_size,
           HH_exp, no_of_earners, HH_size, Percent_Income_Spent_on_Housing, City)
)
# Align and add to summary dataset
rw_final_combined_aligned <- align_columns(rw_final_combined, common_columns)
summary_dataset <- bind_rows(summary_dataset, rw_final_combined_aligned)
# combined_sample_size <- bind_rows(combined_sample_size, sample_size_rw,
sample_size_rw_u_r)
```

```
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"))
```

In addition to the quantile-level summaries, we also generated **aggregated metrics at both state and national levels** to support broader comparisons within and across countries:

- **State-Level Aggregation:**
  The dataset was grouped by Country, Location (region), Year, Quantile, and urban_rural. Within each group, we calculated average household expenditure, household size, and percentage of income spent on housing. The sample sizes were summed to reflect total representation. Kenya was excluded from this step as it is handled separately due to unique formatting.
- **National-Level Aggregation:**
  All regions within each country were collapsed into a national summary by grouping on Country, Year, Quantile, and urban_rural. This allowed for country-level benchmarking across time and income segments.
- **Combined Dataset Creation:**
  The original quantile-level data, state-level aggregates, and national-level aggregates were combined using bind_rows() into a single unified dataset.

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

Calculating Income Growth

All household income data for Rwanda 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 5.896% was used.

Year	GDP per capita growth rate (%)
2000	7.499871
2001	7.526799
2002	11.66824
2003	-0.01969
2004	4.645119
2005	6.506697
2006	6.352667
2007	4.80791
2008	8.277574
2009	3.542919
2010	4.659798
2011	5.342777
2012	6.071727
2013	2.233015
2014	3.631646
2015	6.282592
2016	3.488788
2017	1.509065
2018	6.03178
2019	6.979296
2020	-5.51637
2021	8.456009
2022	5.817733
2023	5.896001
2024	6.583267
2025	5.896001

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