Assignement 2

2024-01-10

Introduction :

This assignment consists of performing data cleaning and manipulation, and then some statistical analysis.

Dataset:

The dataset is retrieved from Rwanda DHS (Demographic and Health Survey) 2020. The type of dataset used here is Household member. You will get data in two files: main SPSS File and Map File (for descriptions).

Your Assignments steps:

1. Read the dataset in R.

The dataset has 55920 observations and 581 features

- Visualize, inspect and get familiar with the data

```
## # A tibble: 6 x 581
##   HHID       HVIDX HV000 HV001 HV002 HV003 HV004  HV005 HV006 HV007 HV008 HV008A
##   <chr>      <dbl> <chr> <dbl> <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl>  <dbl>
## 1 "       1~     1 RW7       1     1 1         1 1.45e6     6  2020  1446  43994
## 2 "       1~     1 RW7       1     3 1         1 1.45e6     6  2020  1446  43993
## 3 "       1~     2 RW7       1     3 1         1 1.45e6     6  2020  1446  43993
## 4 "       1~     3 RW7       1     3 1         1 1.45e6     6  2020  1446  43993
## 5 "       1~     4 RW7       1     3 1         1 1.45e6     6  2020  1446  43993
## 6 "       1~     5 RW7       1     3 1         1 1.45e6     6  2020  1446  43993
## # i 569 more variables: HV009 <dbl>, HV010 <dbl>, HV011 <dbl>, HV012 <dbl>,
## #   HV013 <dbl>, HV014 <dbl>, HV015 <dbl+lbl>, HV016 <dbl>, HV017 <dbl>,
## #   HV018 <dbl>, HV019 <dbl>, HV020 <dbl+lbl>, HV021 <dbl>, HV022 <dbl+lbl>,
## #   HV023 <dbl+lbl>, HV024 <dbl+lbl>, HV025 <dbl+lbl>, HV026 <dbl+lbl>,
## #   HV027 <dbl+lbl>, HV028 <dbl>, HV030 <dbl>, HV031 <dbl>, HV032 <dbl>,
## #   HV035 <dbl>, HV040 <dbl>, HV041 <dbl>, HV042 <dbl+lbl>, HV044 <dbl+lbl>,
## #   HV045A <dbl+lbl>, HV045B <dbl+lbl>, HV045C <dbl+lbl>, HV046 <dbl+lbl>, ...
```

```
## # A tibble: 6 x 581
##   HHID       HVIDX HV000 HV001 HV002 HV003 HV004  HV005 HV006 HV007 HV008 HV008A
##   <chr>      <dbl> <chr> <dbl> <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl>  <dbl>
## 1 "     500~     2 RW7     500    26 1       500 1.02e6     3  2020  1443  43902
## 2 "     500~     3 RW7     500    26 1       500 1.02e6     3  2020  1443  43902
## 3 "     500~     1 RW7     500    27 2       500 1.02e6     3  2020  1443  43902
## 4 "     500~     2 RW7     500    27 2       500 1.02e6     3  2020  1443  43902
## 5 "     500~     3 RW7     500    27 2       500 1.02e6     3  2020  1443  43902
```

```
## 6 "     500~     4 RW7     500    27 2       500 1.02e6     3  2020  1443  43902
## # i 569 more variables: HV009 <dbl>, HV010 <dbl>, HV011 <dbl>, HV012 <dbl>,
## #   HV013 <dbl>, HV014 <dbl>, HV015 <dbl+lbl>, HV016 <dbl>, HV017 <dbl>,
## #   HV018 <dbl>, HV019 <dbl>, HV020 <dbl+lbl>, HV021 <dbl>, HV022 <dbl+lbl>,
## #   HV023 <dbl+lbl>, HV024 <dbl+lbl>, HV025 <dbl+lbl>, HV026 <dbl+lbl>,
## #   HV027 <dbl+lbl>, HV028 <dbl>, HV030 <dbl>, HV031 <dbl>, HV032 <dbl>,
## #   HV035 <dbl>, HV040 <dbl>, HV041 <dbl>, HV042 <dbl+lbl>, HV044 <dbl+lbl>,
## #   HV045A <dbl+lbl>, HV045B <dbl+lbl>, HV045C <dbl+lbl>, HV046 <dbl+lbl>, ...
```

2. Select only few columns, important in this Assignments. They are the following: "HV001", "HV009", "HV010", "HV011", "HV014", "SHDISTRICT", "HV024", "HV025", "HV040", "HV227", "HV228", "HV270", "HV105", "HV106", "HML3", "HML4", "HML7", "HML10", "HML22", "HML32","HML33", "HML35"

3. Rename variables using the variable descriptions below. Give meaningful (short) name to the variables of your choice.

   - HV001= "Cluster number", -> "cluster_no"
   - HV009 = "Number of household members", -> "no_HH_member"
   - HV010 = "Number of eligible women in household", -> "no_elig_wom_HH"
   - HV011 = "Number of eligible men in household", -> "no_elig_men_HH"
   - HV014 = "Number of children 5 and under (de jure)", -> "child_under_5years"
   - SHDISTRICT = "District (geographic area)", -> "district"
   - HV024 = "Region (provinces, corresponding values in a map file)", -> "province"
   - HV025 = "Type of place of residence (rural versus urban)", -> "locality
   - HV040 = "Cluster altitude in meters", -> "cluster_alt_M"
   - HV227 = "Presence of mosquito bed net for sleeping", -> "mosquito_net_yes"
   - HV228 = "Number of children under 5 who slept under a mosquito bed net", -> "child_under_5years_net_yes
   - HV270 = "Wealth index combined (an index based on various household assets indicating socio-economic status)", -> "socio-economic_status
   - HV105 = "Age of household members", -> "HH_members_ages"
   - HV106 = "Highest educational level attained by individuals", -> "education_level"
   - HML3 = "Net observed by interviewer", -> "net_observed"
   - HML4 = "Months ago the net was obtained", -> "months_since_net_obtained"
   - HML7 = "Brand of net", -> "net_brand"
   - HML10 = "Insecticide-Treated Net (ITN)", -> "ITN"
   - HML22 = "Obtained net from campaign, antenatal, or immunization visit", ->
   - HML33= "Result of malaria measurement", -> "malaria_measures"
   - HML32 = "Final result of malaria from blood smear test", -> "malaria_blood_T_result"
   - HML35 = "Result of malaria rapid test" -> "malaria_rapid_T_result"

Data cleaning

1. Inspect each variables, decode variable to its original unique variables. Example, Variable "HV024"(Region) has Unique values 1,2,3,4,5. Decode it to orginal Region Kigali, South, West, North, East Use Map file to see the description of each values in data.

```
##   cluster_no no_HH_member no_elig_wom_HH no_elig_men_HH child_under_5years
## 1    numeric      numeric        numeric        numeric            numeric
## 2    numeric      numeric        numeric        numeric            numeric
## 3    numeric      numeric        numeric        numeric            numeric
##         district       province       locality cluster_alt_M mosquito_net_yes
```

```
## 1 haven_labelled haven_labelled haven_labelled       numeric   haven_labelled
## 2     vctrs_vctr     vctrs_vctr     vctrs_vctr       numeric       vctrs_vctr
## 3         double         double         double       numeric           double
##   child_under_5years_net_yes socio.economic_status HH_members_ages
## 1             haven_labelled        haven_labelled  haven_labelled
## 2                 vctrs_vctr            vctrs_vctr      vctrs_vctr
## 3                     double                double          double
##   education_level   net_observed months_since_net_obtained      net_brand
## 1  haven_labelled haven_labelled            haven_labelled haven_labelled
## 2      vctrs_vctr     vctrs_vctr                vctrs_vctr     vctrs_vctr
## 3          double         double                    double         double
##              ITN      antenatal malaria_blood_T_result malaria_measures
## 1 haven_labelled haven_labelled         haven_labelled   haven_labelled
## 2     vctrs_vctr     vctrs_vctr             vctrs_vctr       vctrs_vctr
## 3         double         double                 double           double
##   malaria_rapid_T_result
## 1         haven_labelled
## 2             vctrs_vctr
## 3                 double
```

Upon examining the column classes, we discovered that some were labeled as `haven_labelled`. Referring to the **map file**, we chose to decode these columns to their original values using the `as_factor` method.

2. Handling Missing Values:

Determine columns with missing values. Devise the strategy to handle missing values: Deleting missing values, replacing missing values with mean or mode.

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

##                             null_count null_proprition
## malaria_blood_T_result           44869          80.24%
## malaria_rapid_T_result           44851          80.21%
## malaria_measures                 44830          80.17%
## net_observed                     29215          52.24%
## months_since_net_obtained        29215          52.24%
## net_brand                        29215          52.24%
## ITN                              29215          52.24%
## antenatal                        29215          52.24%
## child_under_5years_net_yes       23924          42.78%
## cluster_no                           0             0%
## no_HH_member                         0             0%
## no_elig_wom_HH                       0             0%
```

```
## no_elig_men_HH                    0             0%
## child_under_5years                0             0%
## district                          0             0%
## province                          0             0%
## locality                          0             0%
## cluster_alt_M                     0             0%
## mosquito_net_yes                  0             0%
## socio-economic_status             0             0%
## HH_members_ages                   0             0%
## education_level                   0             0%
```

From the table above, it is evident that some columns have a significant proportion of missing values, ranging from **42%** to **80%**. Typically, columns with **80%** missing data would be discarded. However, as these columns are critical to our analysis, we opted to impute the missing values. For numerical columns, we used the **mean**, while for categorical columns, we used the **mode**.

```
##                            null_count null_proprition
## cluster_no                          0              0%
## no_HH_member                        0              0%
## no_elig_wom_HH                      0              0%
## no_elig_men_HH                      0              0%
## child_under_5years                  0              0%
## district                            0              0%
## province                            0              0%
## locality                            0              0%
## cluster_alt_M                       0              0%
## mosquito_net_yes                    0              0%
## child_under_5years_net_yes          0              0%
## socio-economic_status               0              0%
## HH_members_ages                     0              0%
## education_level                     0              0%
## net_observed                        0              0%
## months_since_net_obtained           0              0%
## net_brand                           0              0%
## ITN                                 0              0%
## antenatal                           0              0%
## malaria_blood_T_result              0              0%
## malaria_measures                    0              0%
## malaria_rapid_T_result              0              0%
```

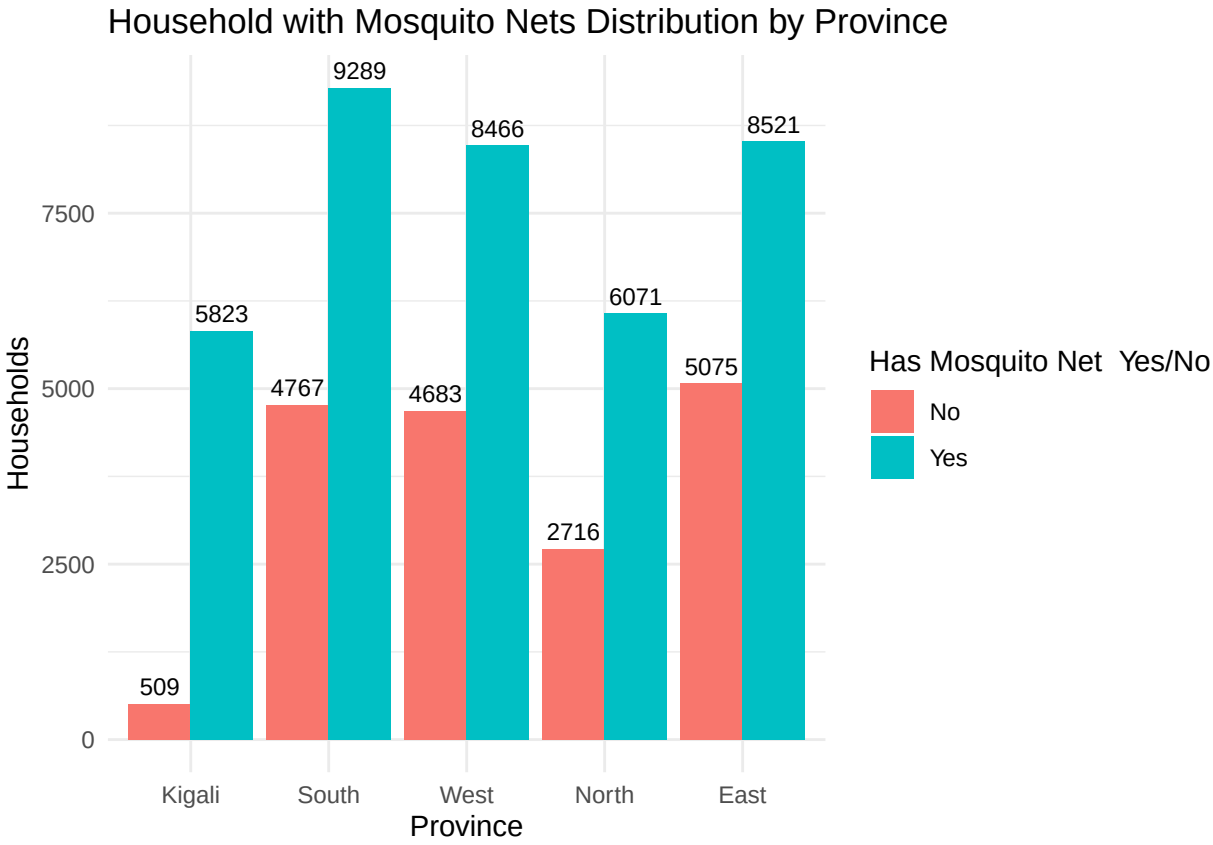
After imputation, all columns are now free of `missing values`.

3. Create new variables

    a. Create variable called "Old Mosquito" variable HML4 (Months ago the net was obtained). The created variable must binary with 1 when mosquito is more than 24 months old.
    b. Create Variable "Average District altitude". Create this variable by averaging cluster altitude in each district. We have three variables HV001= "Cluster number", SHDISTRICT = "District (geographic area)" and HV040 = "Cluster altitude in meters". Filter out clusters in each district, do `mean` of cluster altitude in that district.

Data visualizations:

Produce visualization of your choice. At least each of these - Bar plot

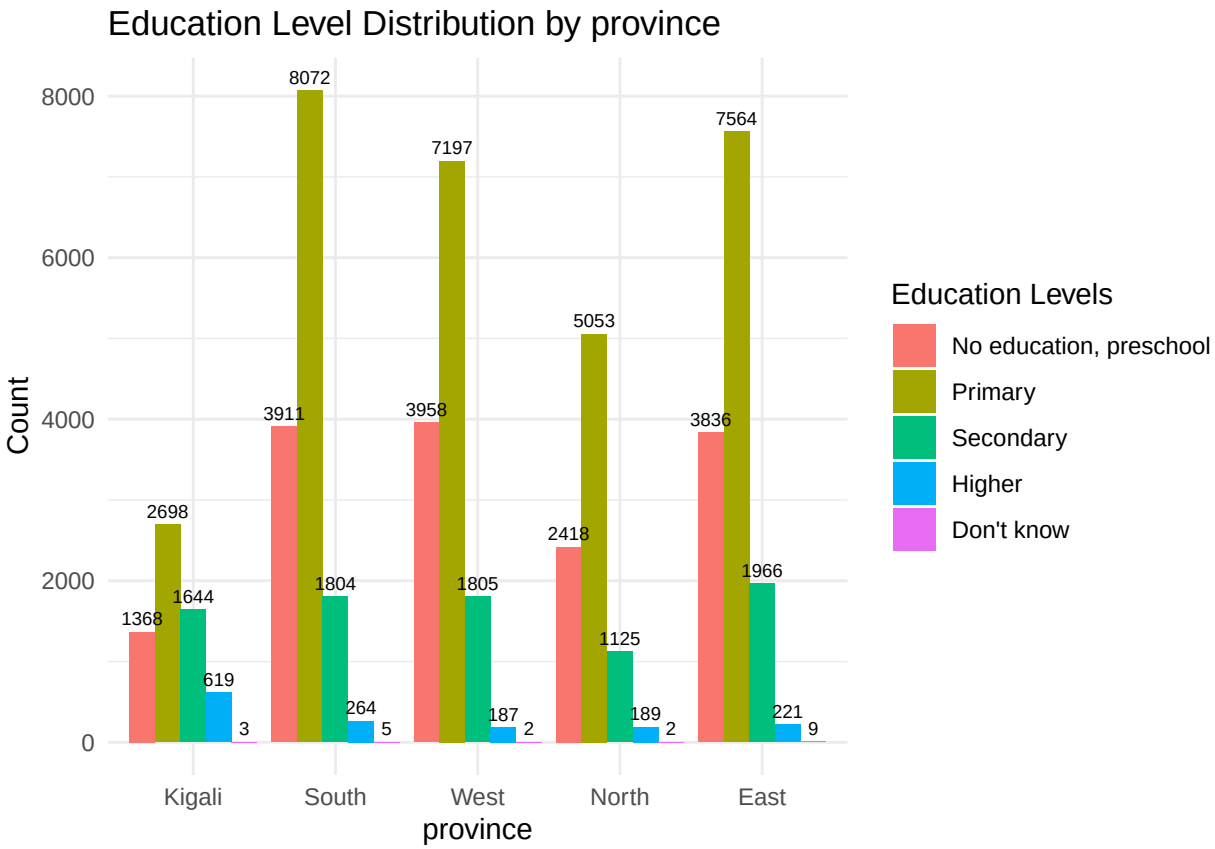


1.**Regional Distribution**

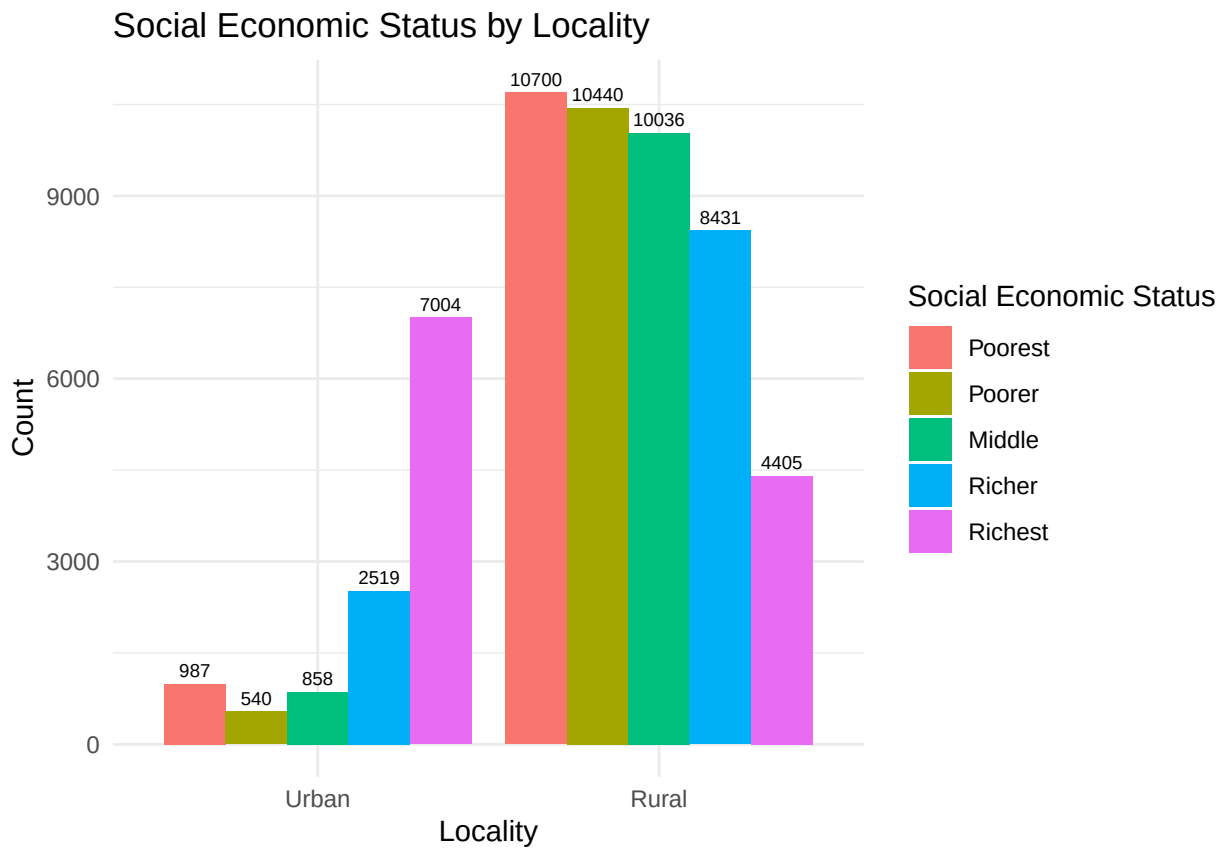
The **East Province** has the highest number of households with mosquito nets (`8,521`), followed closely by the ***South Province** (`9,289`). The **North Province** has a relatively lower number of households with mosquito nets compared to other provinces.

2.**Urban vs Rural Trends**

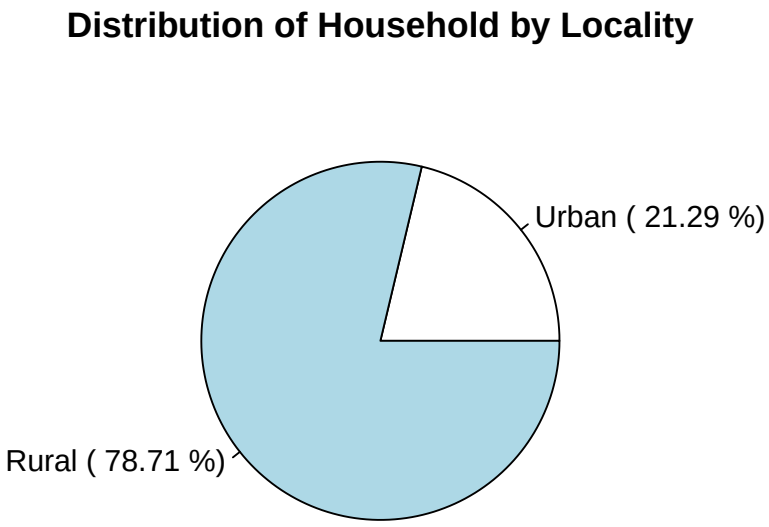
In **Kigali**, a largely urban area, there is a stark contrast, with very few households (509) without mosquito nets. This could indicate better coverage programs or accessibility in urban settings.



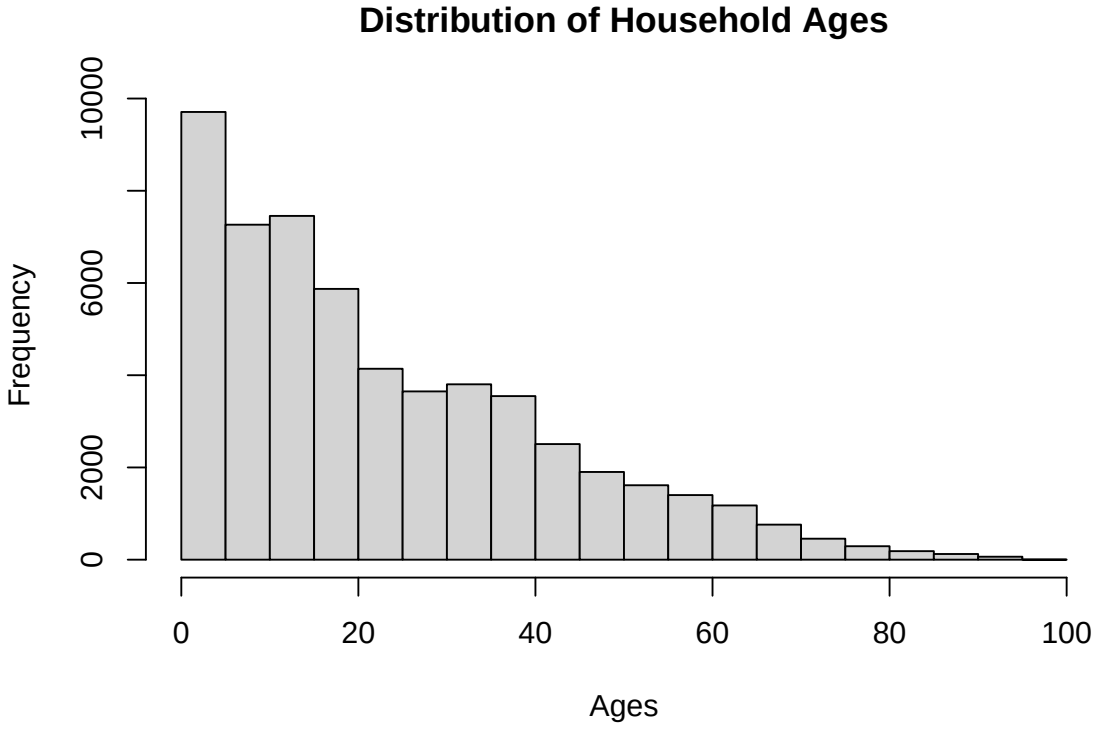
The graph highlights that `primary education` dominates across all provinces, with the **South** (8,072) and **East** (7,564) provinces leading in numbers, reflecting a strong emphasis on basic education. **Kigali** stands out with a relatively higher number of individuals with `higher education` (619), showcasing its urban advantage, while other provinces (North, West, South, and East) lag significantly, with higher education counts below 300, indicating limited access to advanced education. The **North** (5,053) and **East** (3,836) provinces have a notable number of individuals with `no education` or `preschool-level` education. Lastly, **Don't Know** responses are minimal, suggesting respondents whom we consider as those did not want to give the information.



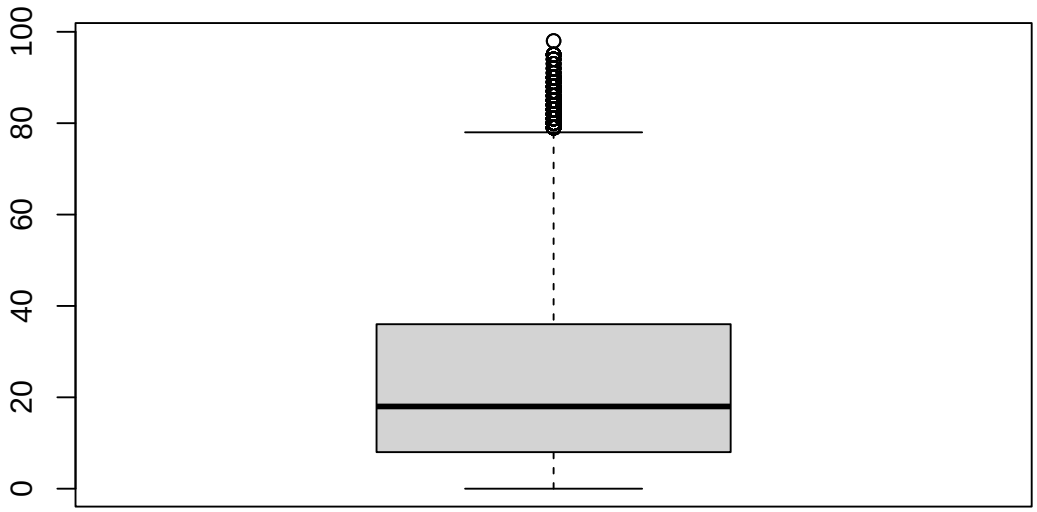
The **rural population** is predominantly concentrated in the `poorest (10,700)` and `poorer (10,440)` categories, highlighting widespread poverty in these areas. In contrast, **urban localities** have a significantly higher representation in the '`richest category (7,004)`. The `middle` and `richer` categories also show a stronger presence in **rural areas**, though the **urban population** contributes a smaller yet notable portion to these categories.



- Pie plot
  The majority of households in survey are in rural areas. 78.71% of households are located in rural areas, while only 21.29% are in urban areas.



- Histogram



- Boxplot

Statistical analysis

Descriptive statistics

1. Use Variable "HML33" to filter out people who had Malaria measurement.

```
## Measured
##    55883
```

2. Calculate Malaria Prevalence for both "Blood Smear" and "Rapid Test"

```
##
##   Negative   Positive No present    Refused      Other
##      55707        185          8         17          3

## Positive
##     0.13

## Positive
##     0.33
```

3. Aggregate Prevalence at district Level

```
## # A tibble: 30 x 3
## # Groups:   district [30]
##    district RPT_prev BT_prev
##    <fct>       <dbl>   <dbl>
##  1 Gasabo       0.71    0.22
##  2 Gisagara     0       0.06
##  3 Gakenke      0       0
##  4 Kayonza      0.69    0.3
##  5 Gatsibo      0.11    0
##  6 Rutsiro      0       0
##  7 Karongi      0.06    0.11
##  8 Muhanga      0.59    0.12
##  9 Gicumbi      0.91    0.23
## 10 Musanze      0.11    0
## # i 20 more rows
```

Analytical Analysis

1. Compare the prevalence in both tests and state if they are different.

Hint: Check `?` the documentations for `t.test` and `aov`.

```
##
##  Paired t-test
##
## data:  district_pre$malaria_rapid_prevalence and district_pre$malaria_blood_prevalence
## t = 3.8637, df = 29, p-value = 0.0005785
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  0.09664123 0.31402544
## sample estimates:
## mean difference
##       0.2053333
```

The paired t-test results show a statistically significant difference between Blood Smear and Rapid Test measurements ($t = -3.8688$, $p = 0.00057$, $df = 29$), indicating that the mean of Rapid Test is higher than that of Blood Smear by approximately -0.206. The 95% confidence interval $[-0.3149, -0.0971]$ supports this conclusion, as it does not include 0. This suggests that Blood Smear consistently produces lower values than Rapid Test.

```
## Call:
##    aov(formula = malaria_rapid_prevalence ~ malaria_blood_prevalence,
##     data = district_pre)
##
## Terms:
##                 malaria_blood_prevalence Residuals
## Sum of Squares                  3.078282  1.898015
## Deg. of Freedom                        1        28
##
## Residual standard error: 0.2603579
## Estimated effects may be unbalanced
```

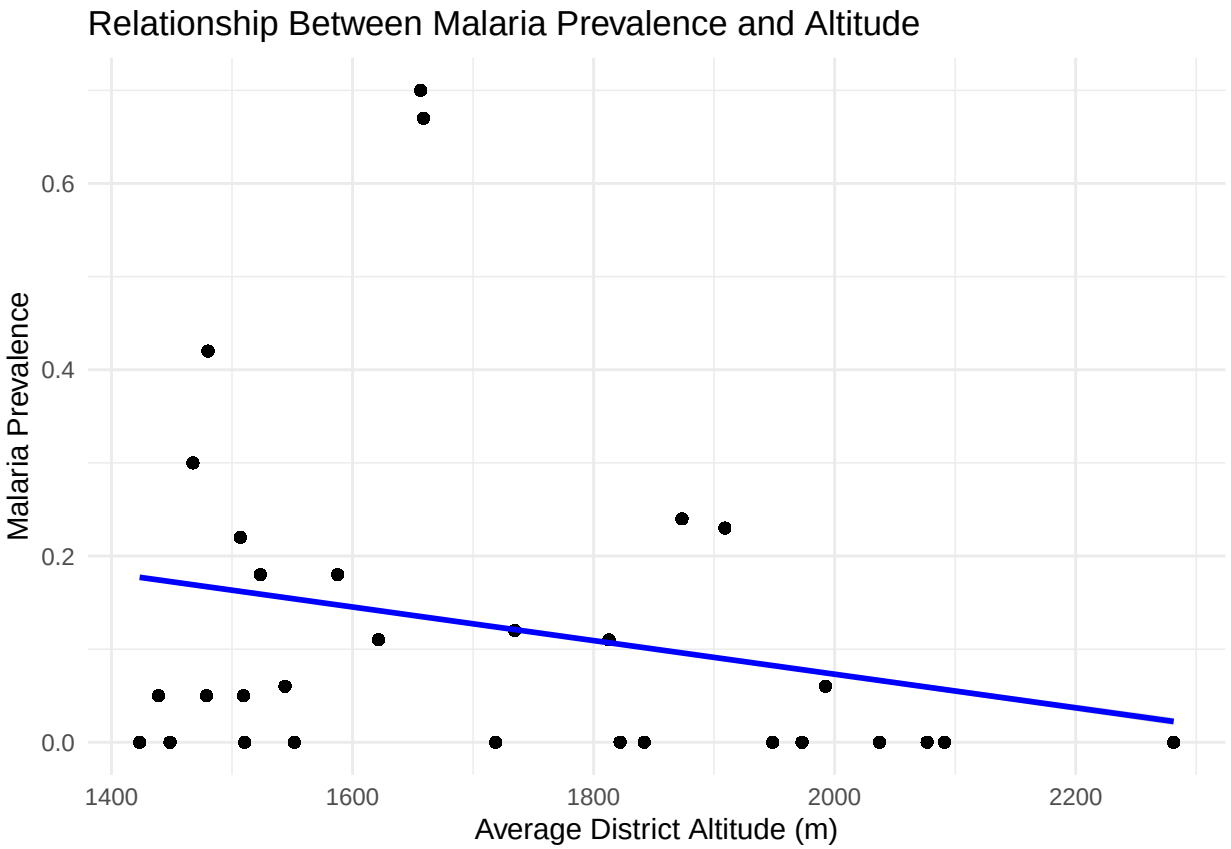
The ANOVA results indicate that the predictor variable has a significant effect on the outcomes ($F = 9.593$, $p = 0.00301$). The degrees of freedom for the test and residuals are 1 and 58, respectively, with the predictor explaining a variability of 0.976 (Sum Sq) compared to 5.900 for the residuals. The mean square for the predictor (0.9760) is notably higher than for the residuals (0.1017), and the highly significant p-value confirms that the predictor meaningfully influences the results. Together with the paired t-test, these findings show that the choice of test method significantly affects outcomes, with the paired t-test focusing on mean differences and ANOVA examining the overall variability.

Bonus

2. Using a statistical model of your choice, determine if there is a relationship between malaria prevalence in a district and its average altitude.

```
##
## Call:
## lm(formula = BT_prev ~ `Average District Altitude`, data = dd)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.17714 -0.11160 -0.05928  0.03249  0.56493
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  4.339e-01  5.560e-03   78.05   <2e-16 ***
## `Average District Altitude` -1.804e-04  3.222e-06  -55.99   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1799 on 55918 degrees of freedom
## Multiple R-squared:  0.05308,    Adjusted R-squared:  0.05307
## F-statistic:  3135 on 1 and 55918 DF,  p-value: < 2.2e-16

## `geom_smooth()` using formula = 'y ~ x'
```



There is a statistically significant relationship between district altitude and malaria prevalence. Higher altitudes are associated with lower malaria prevalence. However, the R-squared value indicates that altitude alone explains a small portion of the variability in malaria prevalence, suggesting other factors also play a significant role.