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.

```
## Warning: package 'haven' was built under R version 4.4.2
```

```
## [1] 55920   581
```

- Visualize, inspect and get familiar with the data

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"

```
## # A tibble: 55,920 x 22
##    HV001 HV009 HV010 HV011 HV014 SHDISTRICT  HV024      HV025     HV040 HV227
##    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl+lbl>   <dbl+lbl>  <dbl+lbl> <dbl> <dbl+lb>
##  1     1     1     0     0     0 12 [Gasabo] 1 [Kigali] 2 [Rural]  1831 1 [Yes]
##  2     1     5     1     1     0 12 [Gasabo] 1 [Kigali] 2 [Rural]  1831 1 [Yes]
##  3     1     5     1     1     0 12 [Gasabo] 1 [Kigali] 2 [Rural]  1831 1 [Yes]
##  4     1     5     1     1     0 12 [Gasabo] 1 [Kigali] 2 [Rural]  1831 1 [Yes]
##  5     1     5     1     1     0 12 [Gasabo] 1 [Kigali] 2 [Rural]  1831 1 [Yes]
##  6     1     5     1     1     0 12 [Gasabo] 1 [Kigali] 2 [Rural]  1831 1 [Yes]
##  7     1     1     1     0     0 12 [Gasabo] 1 [Kigali] 2 [Rural]  1831 1 [Yes]
##  8     1     4     1     1     1 12 [Gasabo] 1 [Kigali] 2 [Rural]  1831 1 [Yes]
##  9     1     4     1     1     1 12 [Gasabo] 1 [Kigali] 2 [Rural]  1831 1 [Yes]
## 10     1     4     1     1     1 12 [Gasabo] 1 [Kigali] 2 [Rural]  1831 1 [Yes]
## # i 55,910 more rows
## # i 12 more variables: HV228 <dbl+lbl>, HV270 <dbl+lbl>, HV105 <dbl+lbl>,
## #   HV106 <dbl+lbl>, HML3 <dbl+lbl>, HML4 <dbl+lbl>, HML7 <dbl+lbl>,
## #   HML10 <dbl+lbl>, HML22 <dbl+lbl>, HML32 <dbl+lbl>, HML33 <dbl+lbl>,
## #   HML35 <dbl+lbl>
```

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

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

Renaming the dataset variables with meaningful names

```
## # A tibble: 6 x 22
##   ClusterNumber HouseholdMembers EligibleWomen EligibleMen ChildrenUnder5
##           <dbl>            <dbl>         <dbl>       <dbl>          <dbl>
## 1             1                1             0           0              0
## 2             1                5             1           1              0
## 3             1                5             1           1              0
## 4             1                5             1           1              0
## 5             1                5             1           1              0
## 6             1                5             1           1              0
## # i 17 more variables: District <dbl+lbl>, Region <dbl+lbl>,
## #   ResidenceType <dbl+lbl>, Altitude <dbl>, BedNetPresence <dbl+lbl>,
## #   ChildrenUnder5BedNet <dbl+lbl>, WealthIndex <dbl+lbl>, Age <dbl+lbl>,
## #   EducationLevel <dbl+lbl>, NetObserved <dbl+lbl>,
## #   NetObtainedMonthsAgo <dbl+lbl>, NetBrand <dbl+lbl>, ITN <dbl+lbl>,
## #   NetSource <dbl+lbl>, MalariaMeasurementResult <dbl+lbl>,
## #   BloodSmearMalariaResult <dbl+lbl>, RapidTestMalariaResult <dbl+lbl>
```

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.

```
# your code
```

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.

```
# your code
```

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

```
# your code
```

- Pie plot

```
# your code
```

- Histogram

```
# your code
```

- Boxplot

```
# your code
```

Statistical analysis

Descriptive statistics

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

```
# your code
```

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

```
# your code
```

3. Aggregate Prevalence at district Level

```
# your code
```

Analytical Analysis

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

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

```
# your code
```

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.

```
# your code
```