SIMAD UNIVERSITY

SCHOOL OF GRADUATE STUDIES

MASTER OF DATA SCIENCE

DATA VISUALIZATION

APRIL 2023 SEMESTER 2

COURSE PROJECT:

SOMALIA POPULATION DATASET ANALYSIS AND VISUALIZATION

BY: ABDIQANI AHMED ABDULLE

LECTURER: DR. FUAD MIRE

Introduction

This dataset contains information about the population of different regions and districts in Somalia. It includes data on total population, internally displaced persons (IDPs), non-displaced persons, IDP children, IDP men, IDP women, IDP elderly, non-displaced children, non-displaced men, non-displaced women, non-displaced elderly, total children, total men, total women and total elderly.

The dataset covers eighteen regions and 75 districts including: Middle Shabelle-Adan Yabaal; Lower Shabelle-Afgooye; Lower Juba-Afmadow; Gedo-Baardheere; Lower Juba-Badhaadhe; Bay-Baidoa; Awdal-Baki; and Middle Shabelle-Balcad. And so on.

Our goal is to analyze this dataset using R programming language and R Studio IDE.

While we are in this process, we make sure to do these steps that involves data analysis:

Data cleaning: here we make sure the dataset looks great and suitable for further analysis. We remove an wanted columns, rename it, check duplicates and remove, check outliers and remove it. And so on.

Data visualization: here is where we where most of our time to make great visualization for the data. We use ggplot liberary in R programming to visualize different graphs eg. Bie, bar map and so on.

Why I chose this dataset

Choosing a dataset is an important step in any data analysis project, and as a student pursuing a Master's degree in Data Science, I have selected the "IDP Population in Somalia" dataset for my course project on data visualization.

The "IDP Population in Somalia" dataset contains information about the number of Internally Displaced Persons (IDPs) in Somalia, categorized by region and district. The dataset includes demographic information, such as age and gender, and provides a breakdown of the total population, IDPs, and non-displaced individuals. The dataset contains 20 variables, including "Total population", "IDPs". "Non displaced,", “IDPs children”, “Non displaced children” etc.

"I chose this dataset for several reasons. Firstly, the dataset is about Somalia, which is my country of origin and where we live, and I am interested in analyzing data related to it. Secondly, as a student with a keen interest in the field of humanitarian aid, I find this dataset highly relevant. The displacement crisis in Somalia has been ongoing for many years, and accurate data on the number of IDPs is crucial in understanding the humanitarian needs of the affected population. Thirdly, the dataset is well-documented and publicly available on data.humdata.org, making it easy to find resources and literature to support my analysis.

Additionally, the dataset needs some minor improvements for my use case, e.g., dealing with outliers, adding new columns or renaming existing ones, which presents an opportunity for me to practice data cleaning and imputation techniques. The dataset also contains a diverse set of variables, making it an excellent choice for data visualization. I plan to use data visualization techniques to explore the distribution of IDPs across regions and districts, identify any patterns or trends in displacement, and determine the demographic profile of the affected population.

In summary, I selected the "IDP Population in Somalia" dataset because of its relevance to the field of humanitarian aid, its significance in understanding the displacement crisis in Somalia, and its diverse set of variables that make it an excellent choice for data visualization. I am excited about exploring this dataset further and applying various data visualization techniques to gain valuable insights that can be used to improve the response to the displacement crisis in Somalia.

Data dictionary

NO	Variable name	Description
1	State	Name of the state
2	Region	Name of the region
3	District	Name of the district
4	Total_population	Total population in the district
5	IDPs	Number of internally displaced persons (IDPs) in the district
6	Non_displaced	Number of non-displaced persons in the district
7	IDPs_children	Number of IDP children in the district
8	IDPs_men	Number of IDP men in the district
9	IDPs_women	Number of IDP women in the district
10	IDPs_elderly	Number of elderly IDPs in the district
11	Non_disp_children	Number of non-displaced children in the district
12	Non_disp_men	Number of non-displaced men in the district
13	Non_disp_women	Number of non-displaced women in the district
14	Non_disp_elderly	Number of elderly non-displaced persons in the district
15	Total_children	Total number of children in the district
16	Total_men	Total number of men in the district
17	Total_women	Total number of women in the district
18	Total_elderly	Total number of elderly persons in the district

Hypotheses

Our hypothesis in this analyze will be as follows:

1.What is the total population distribution in Somalia?

2. How does the IDP population compare to the non-displaced population in Somalia?

3. What is the total population by state?

4. What is the total population by region?

5. What is the total population by district?

6. What is the IDP and non-displaced population like in states?

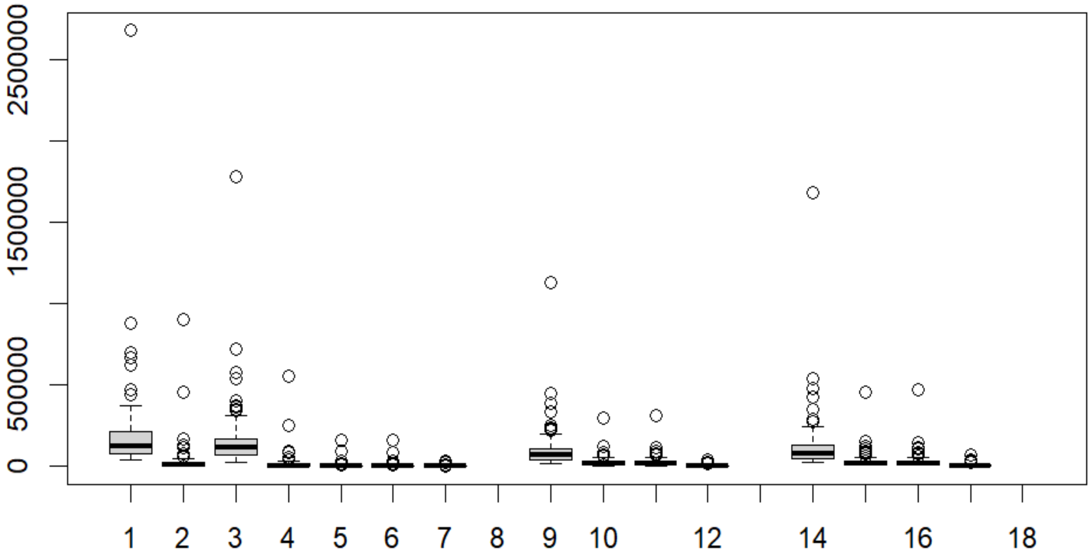
7. What is the distribution of non-displaced population like?

8. What is the distribution of IDPs like in states?

9. How is the population mapped in states?

Data cleaning

Outlier detection and removing



is this boxplot, we can see there is outliers in our dataset that can effect our later analysis. in region columns there is an outlier banaadir who has much more population according to other regions. to solve this we can make a new state column, so states are around same population with banaadir state.

Region statistics:

Mean = 874398.8

Median = 732464

Max = 2683312

Min = 362723

After we made State column and did look statistics results:

Mean = 2248454

Median = 2080632

Max = 2863593

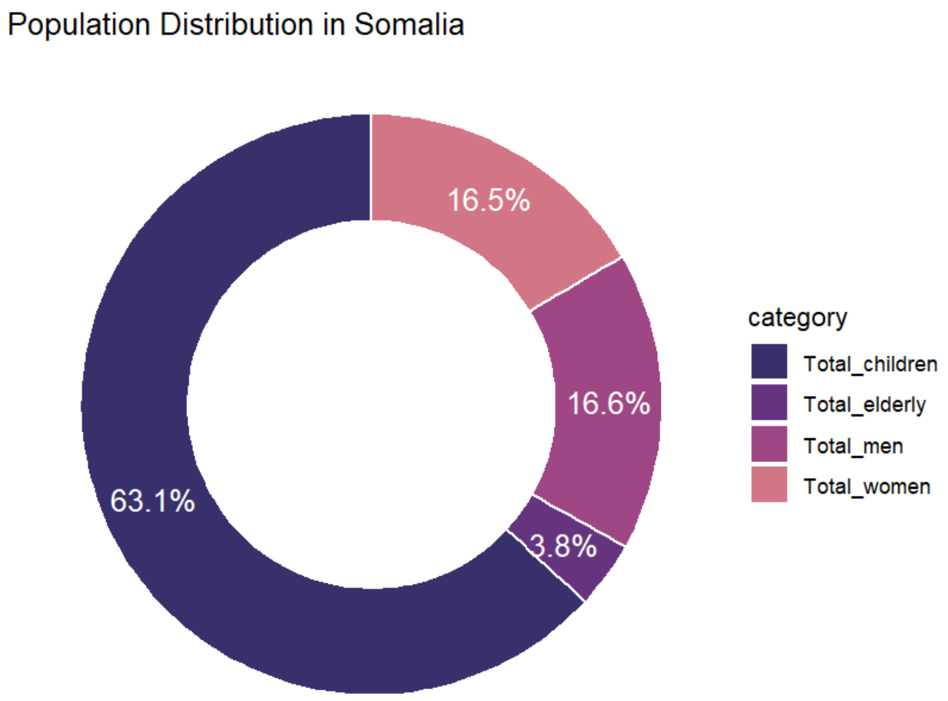
Min = 1284519

This shows how we solved the outlier since Banaadir is much more population than the other regions, to solve this we made a new column called state where we compine regions to their states. Now Banaadir among other states are comparable numbers in population. So, we can go ahead in our analysis and visualizations.

Data analyse and visualization

Here we visualize total population using R ggplot package.

1.What is the total population distribution in Somalia?

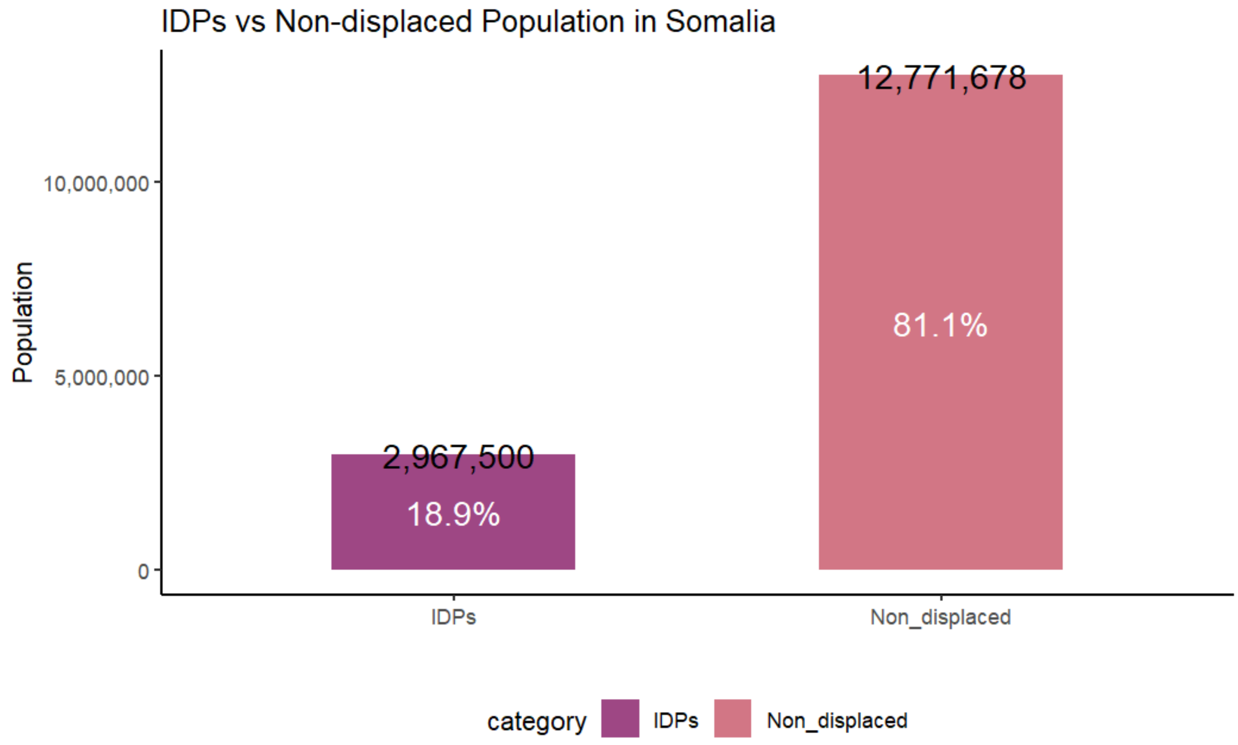


The above bie chart shows population distribution.

Children are 63 percent of the population.

Men and women are almost identically same with 16.5 percent.

2. How does the IDP population compare to the non-displaced population in Somalia?

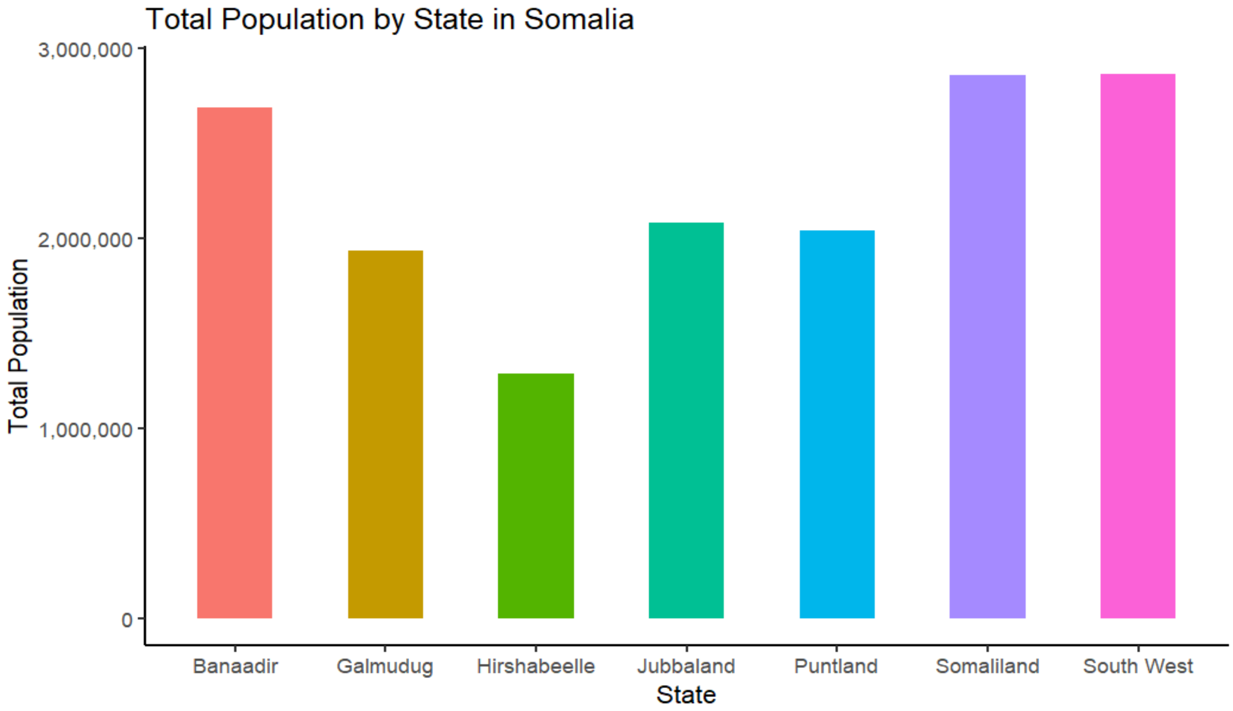


The above bar chart shows IDPs and non displaced population side by side.

81 percent are non displaced population

18 percent are IDPs.

3. What is the total population by state?

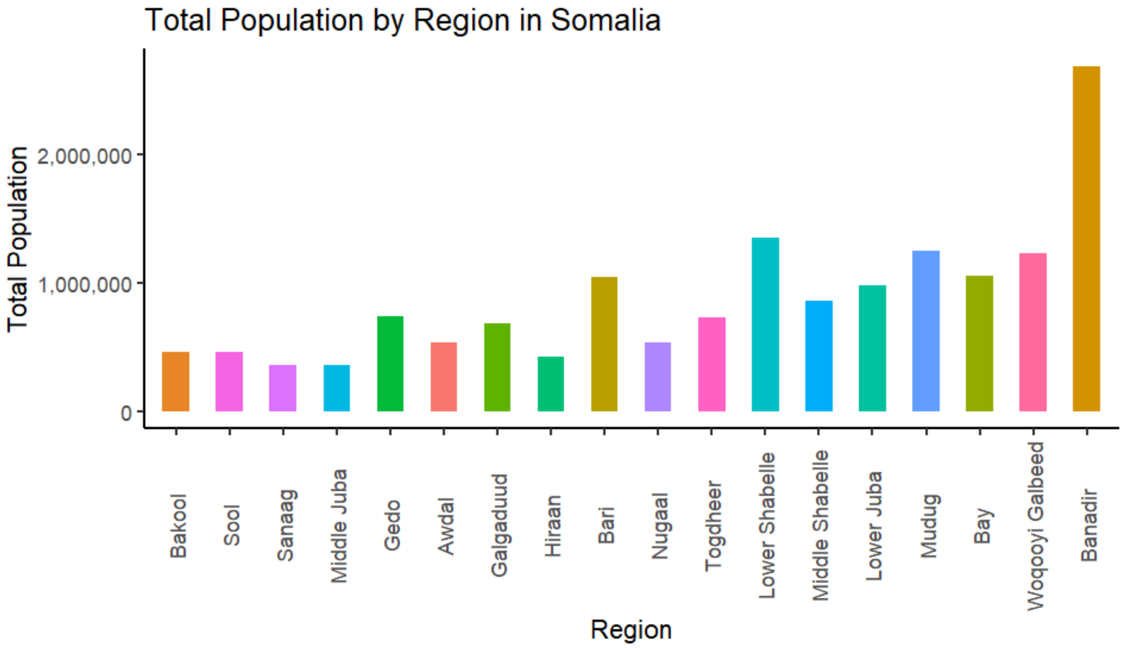


In this bar chart, shows states with total population.

Somaliland and southwest are leading with similar numbers, followed by banaadir and others.

Hirshabeelle is the least population state.

4. What is the total population by region?

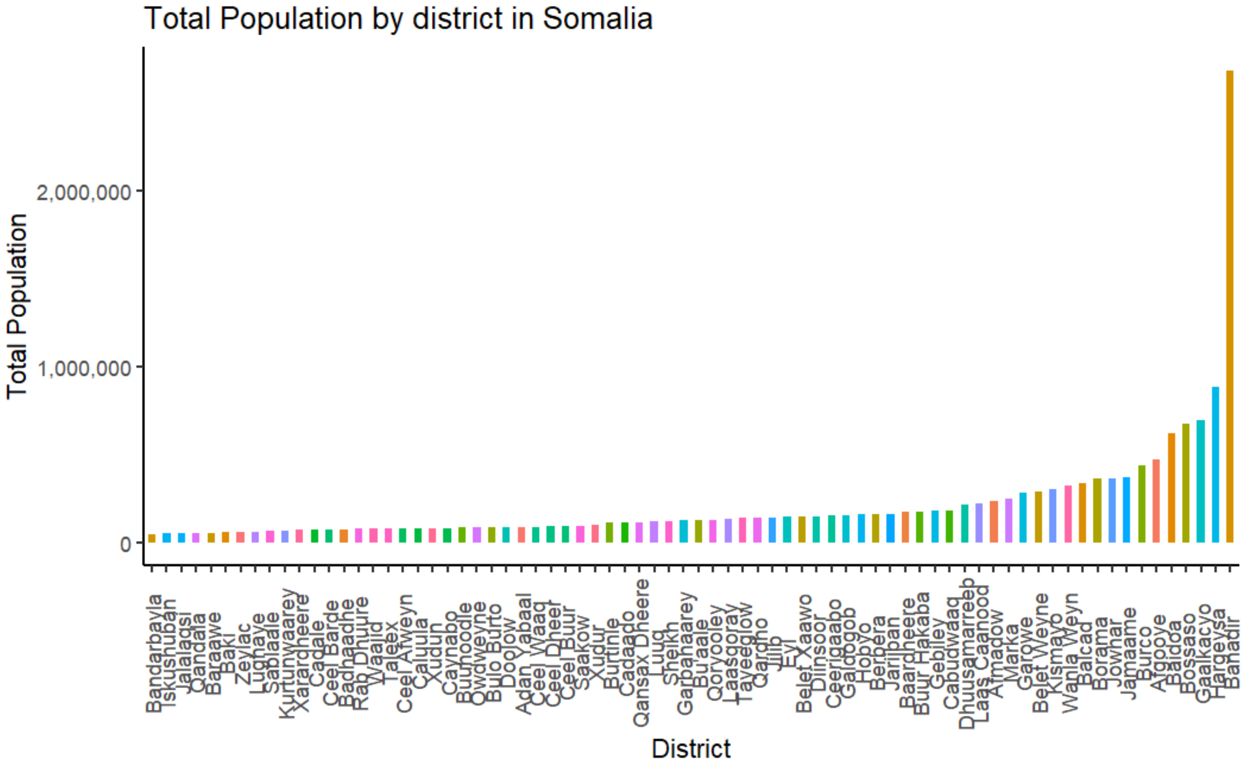


In the above chart, you can see clearly banaadir is the outlier of this view. we solved with state column as shown previous above graph.

Other than Banaadir, lower shabeelle, mudug and waqooyi galbeed are leading the numbers.

Middle juba and sanaag are least populated regions.

5. What is the total population by district?

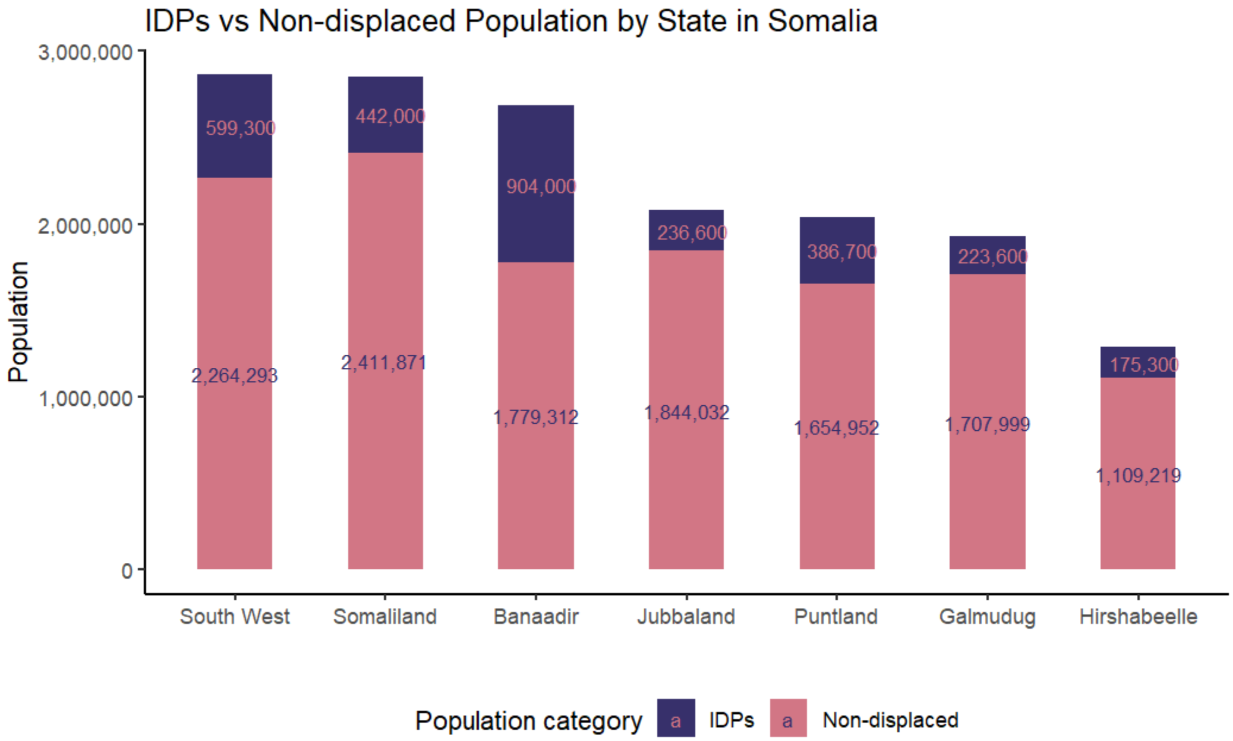


In this chart, shows districts population with banaadir outlier. since banaadir is not listed its districts it is clear that has much population according to other small districts in the country.

Other than Banaadir, hargaisa, gaalkacyo and bossaaso are leading most populated districts.

In the other side, banadarbayla, iskushuban and jalalaqsi are least populated districts.

6. What is the IDP and non-displaced population like in states?

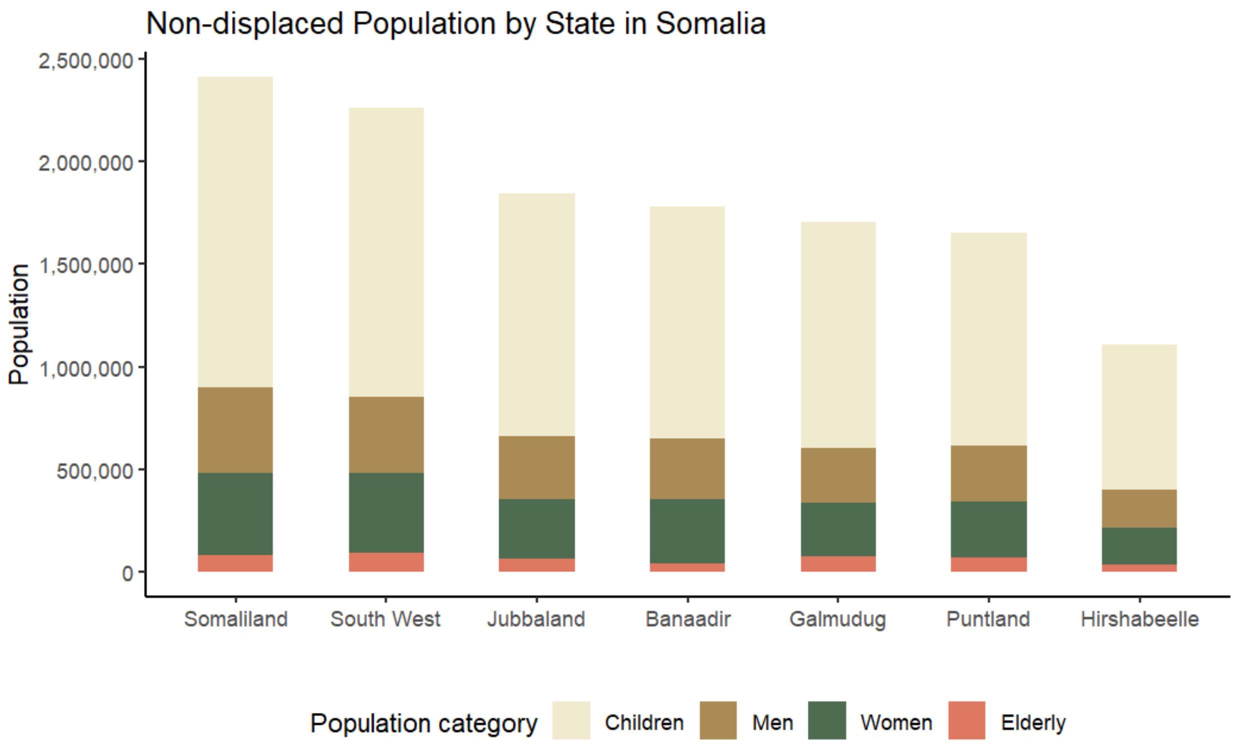


In the above chart, it shows IDPs and non displaced population for every state.

the most IDPs live in banaadir.

the most non displaced people live in somaliland.

7. What is the distribution of non-displaced population like?

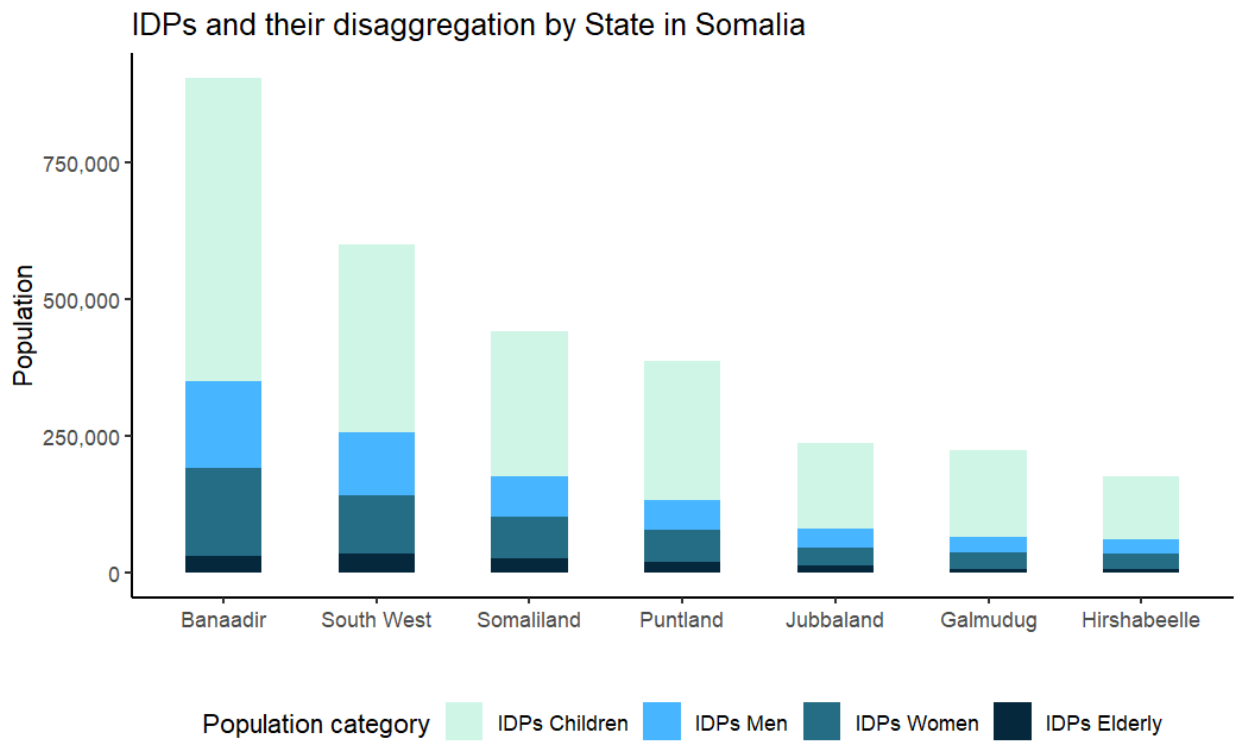


In the above chart, shows non displaced population distribution in states.

Most children and most men live in Somaliland.

Most elderly live in south west.

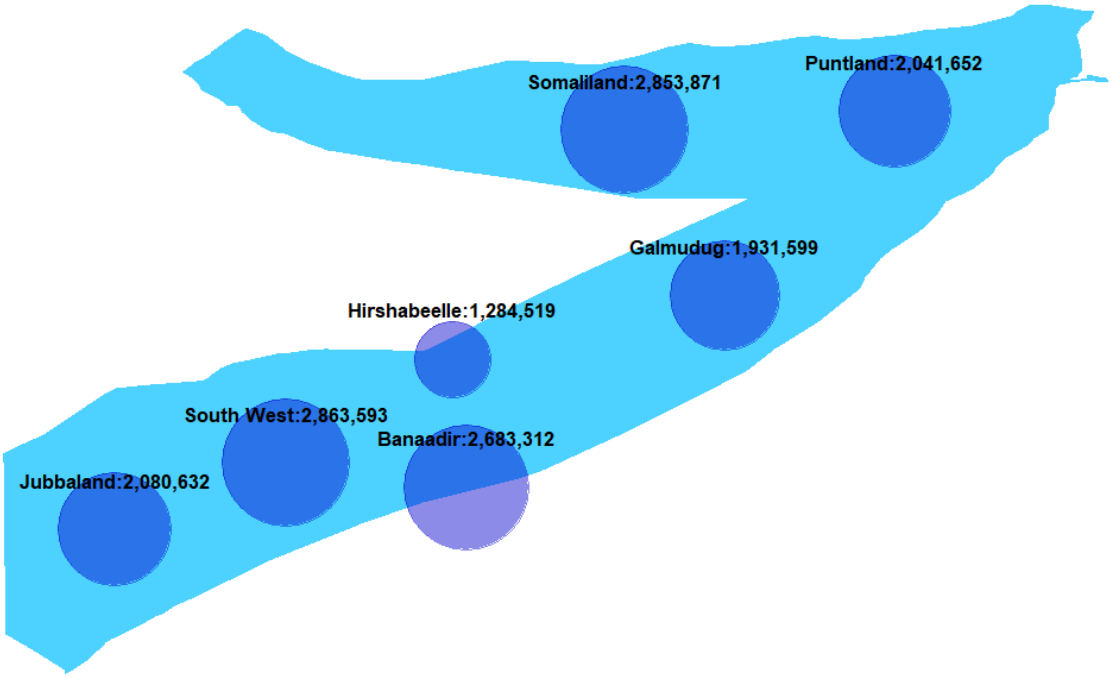
8. What is the distribution of IDPs like in states?



In the above chart, IDPs disagregation by states.

Banaadir are leading every part of disaggregation whether its child, men, women, or elderly. followed by southwest and somaliland.

9. How is the population mapped in states?



The above graph shows, where states are located in the country and total population live in.

we can say that southern area population are more than the northern population.

Conclusion

We completed analyzing and visualizing Somalia population dataset. We made cleaning data process and did the visualization using r packages mostly ggplot and finally we presented to the viewers our insights and finding.