HSNP BENEFICIARY DATA CLEANING

DATA UNDERSTANDING

The dataset contained 570675 rows and 17 columns. Amongst the 17 columns, 9 contained integers. These were:

- Household_ID
- Village_ID
- Sublocation_ID
- Location_ID
- Constituency_ID
- County_ID
- UserCode
- Latitude
- Longitude

There was one datetime column called

- Entry_Date

Also, one column contained Boolean values called

- IsBeneficiaryHH

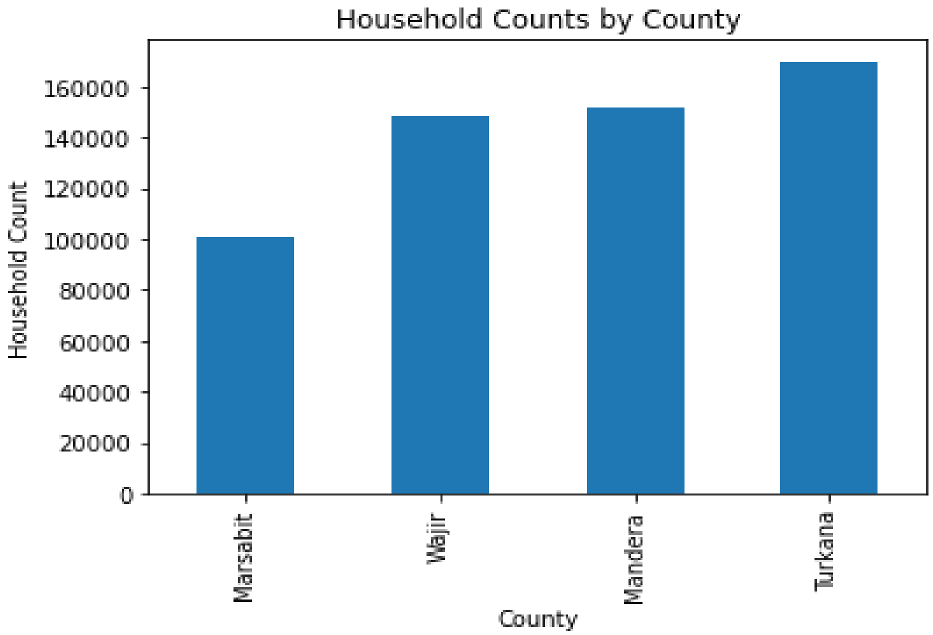
The rest of the columns contained objects.

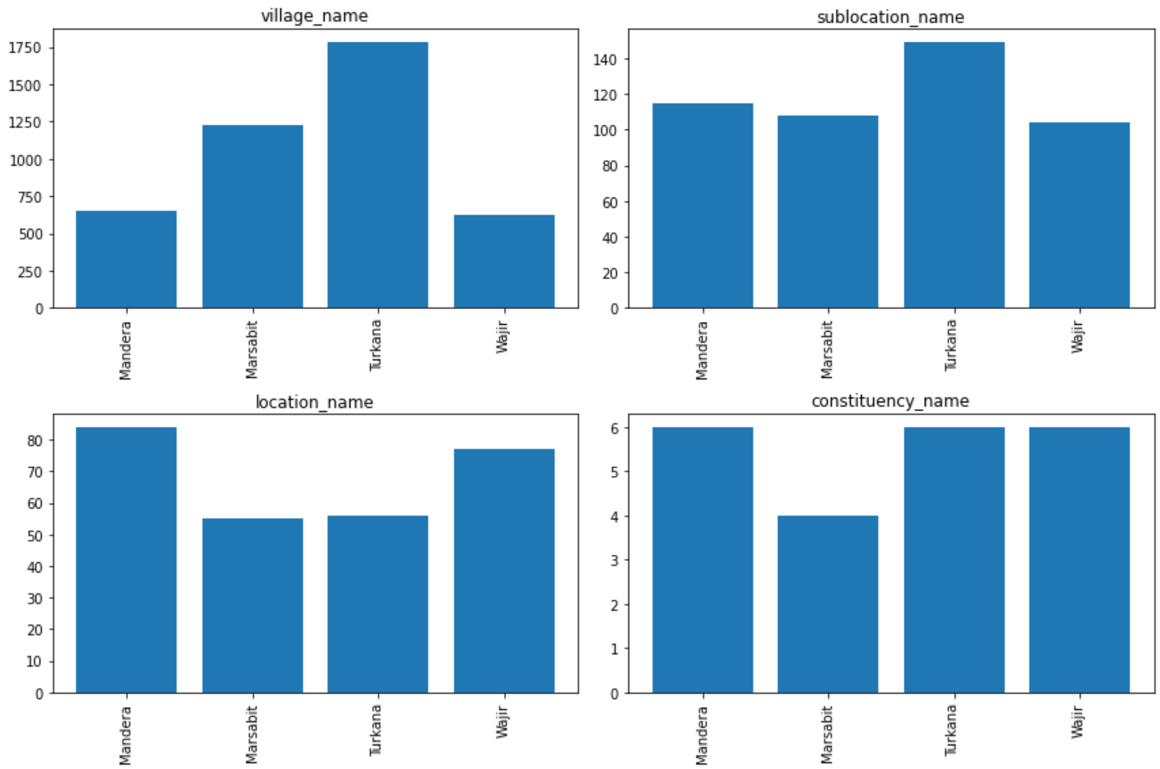
- Village_Name
- Sublocation_Name
- Location_Name
- Constituency_Name
- County_Name
- RuralUrban

Only one column contained missing values which was the ruralurban column. This column contained 26104 missing values which was about 4.5% of the total column.

The dataset contained four counties:

1. Mandera
2. Marsabit
3. Wajir
4. Turkana

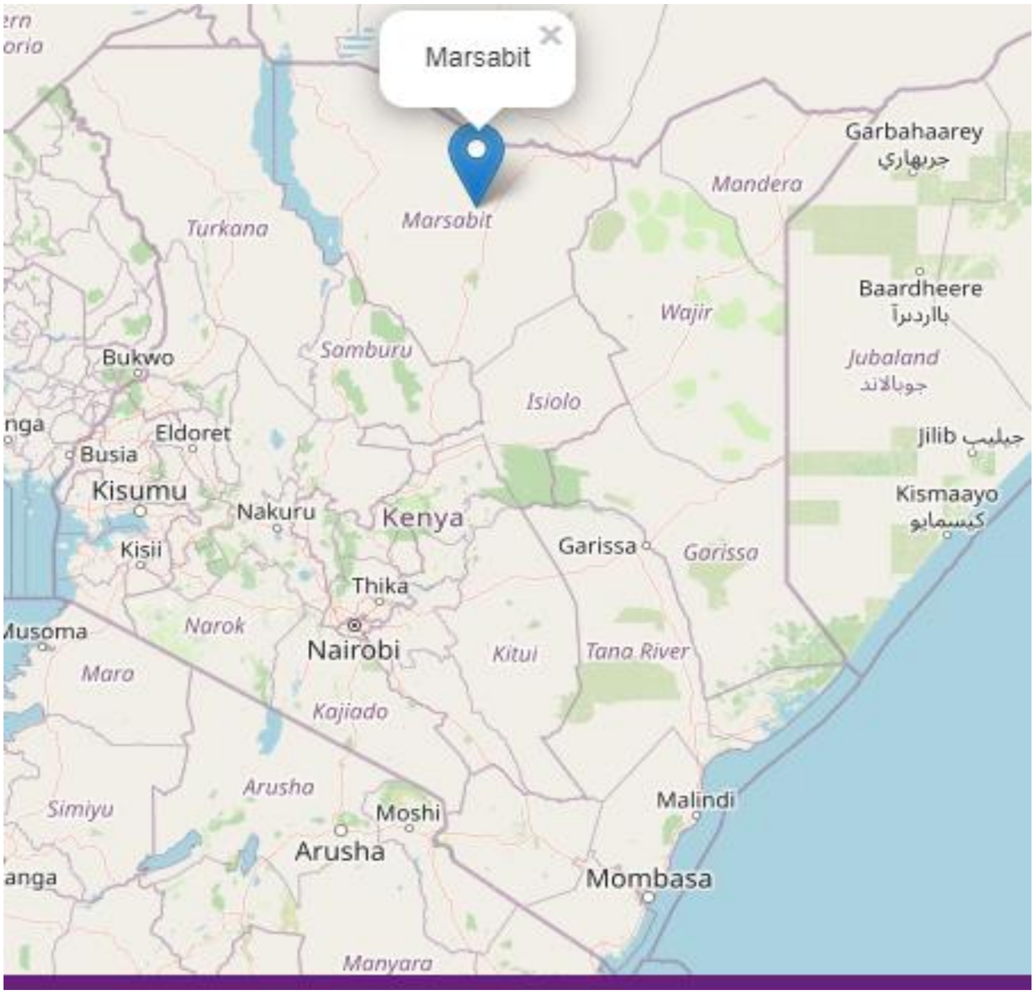



village_name
0
250
500
750
1000
1250
1500
1750
Mandera
Marsabit
Turkana
Wajir
sublocation_name
0
20
40
60
80
100
120
140
Mandera
Marsabit
Turkana
Wajir
location_name
0
10
20
30
40
50
60
70
80
Mandera
Marsabit
Turkana
Wajir
constituency_name
0
1
2
3
4
5
6
Mandera
Marsabit
Turkana
Wajir

DATA CLEANING

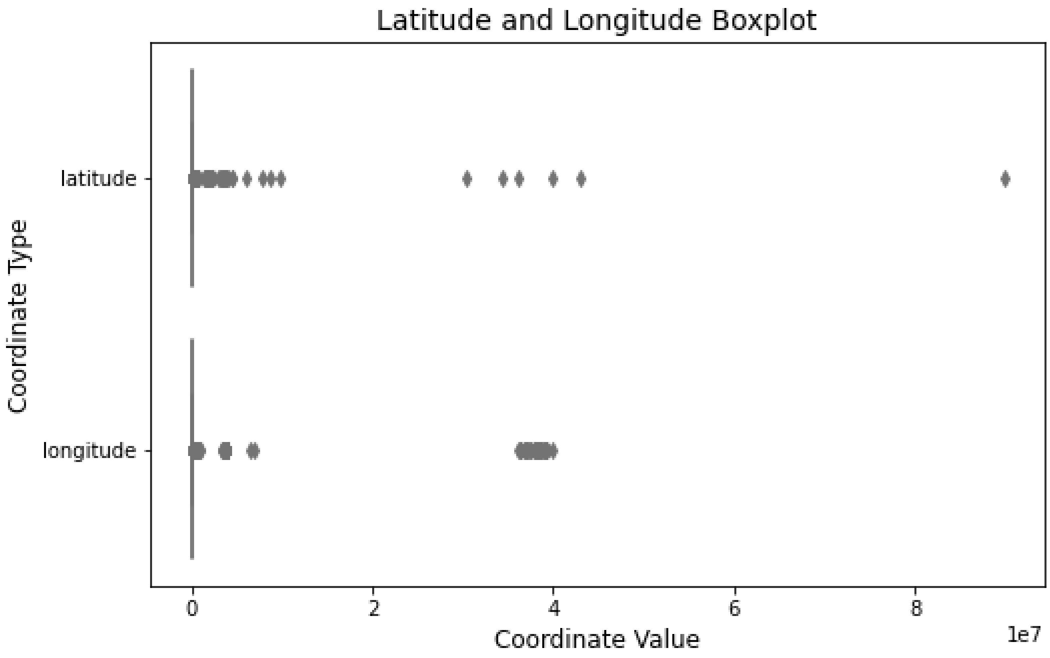
I started off by filtering out each county to enable me clean out the dataset county by county. I used the same methods to clean each county so I'll illustrate using Marsabit county.

1) Marsabit

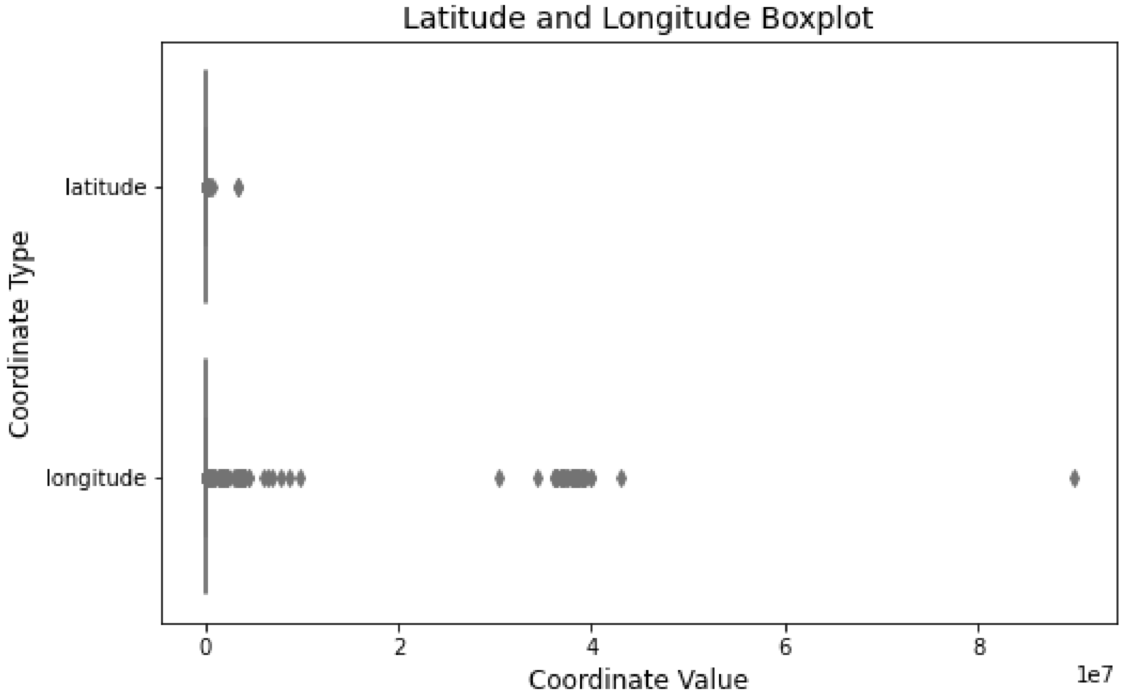


Marsabit contained 100538 beneficiaries.

The latitude and longitude columns contained a lot of outliers.

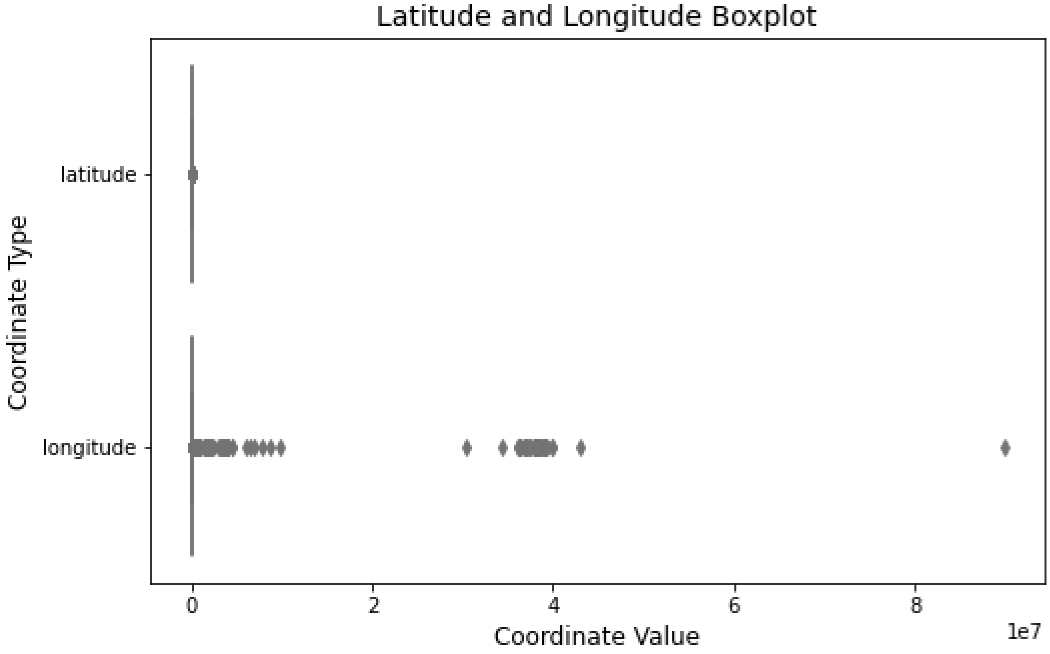


I first started by checking the rows where latitude is greater than longitude in the dataset and swapped their values. This operation helped to ensure that latitude values are always smaller than longitude values, which is the expected convection for geospatial coordinates. After doing this I got the following results

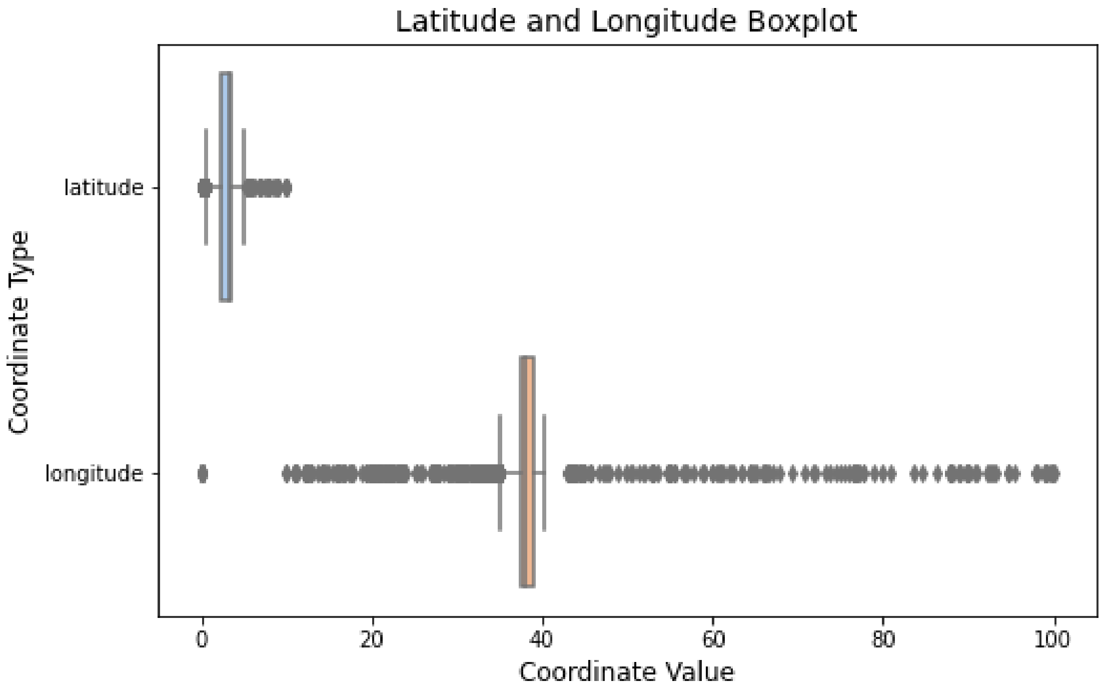


From here, I found the maximum latitude was 3333333.0 and the minimum latitude was 0.0 where as the maximum longitude was 89823455.0 and minimum longitude was 0.0

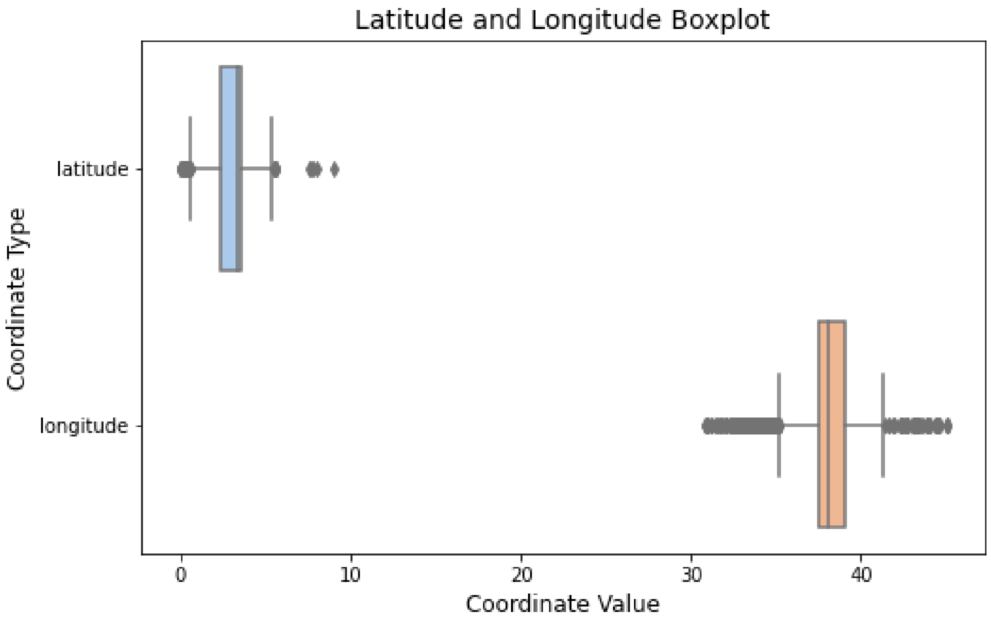
Next, I calculated a correction factor for each latitude value in the dataset based on the length of its integer part. I divided each latitude value by its corresponding correction factor to adjust the values accordingly. After doing that I got the maximum latitude to be 9.9 and minimum latitude still remained at 0.0



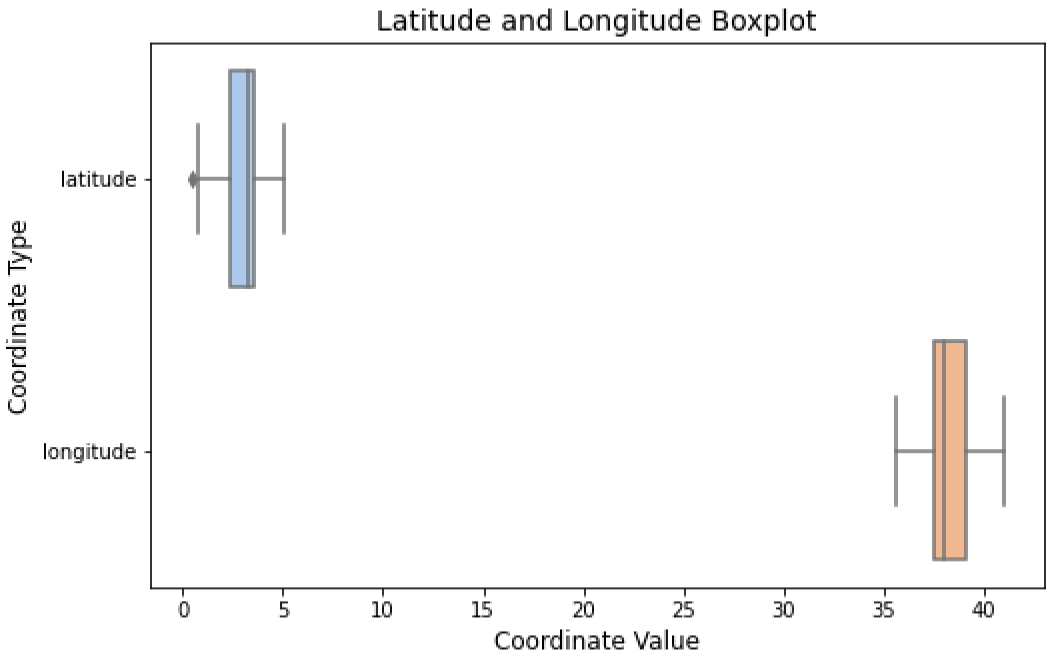
I also calculated a correction factor for each longitude value in the dataset based on the length of its integer part. I divided each longitude value by its corresponding correction factor to adjust the values accordingly. After performing this operation, I was able to adjust the maximum longitude to 99.0 but the minimum longitude remained at 0.0.

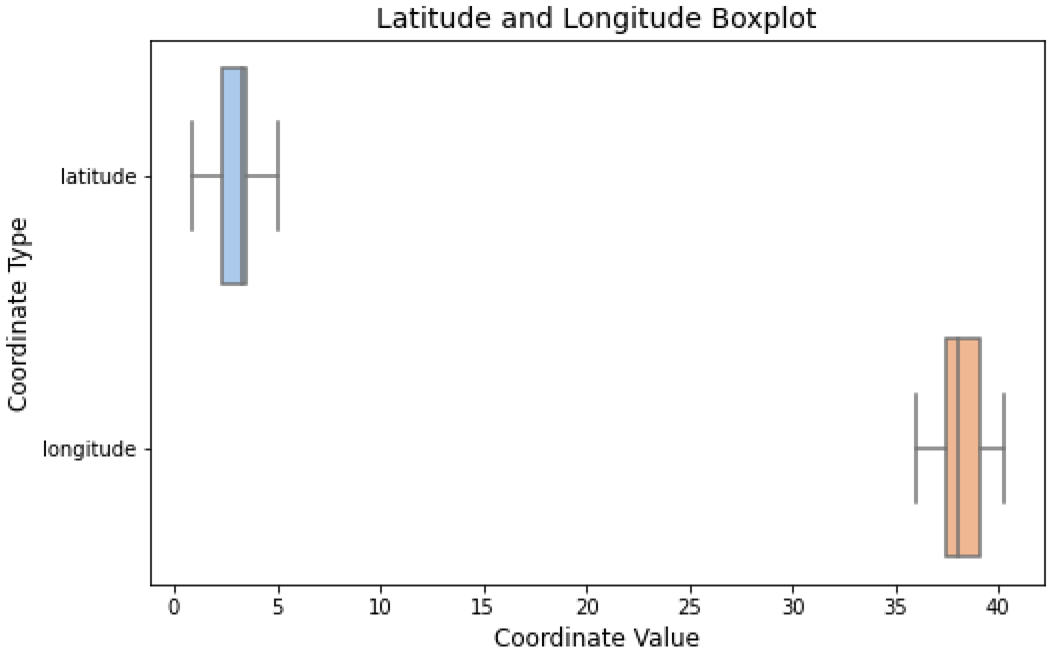


Next, I calculated the mean longitude and latitude values in the dataset. By doing this I was able to identify outliers based on a threshold and replaced those outliers with random values drawn from a normal distribution around the mean. This process helped to mitigate the influence of outliers on the dataset.



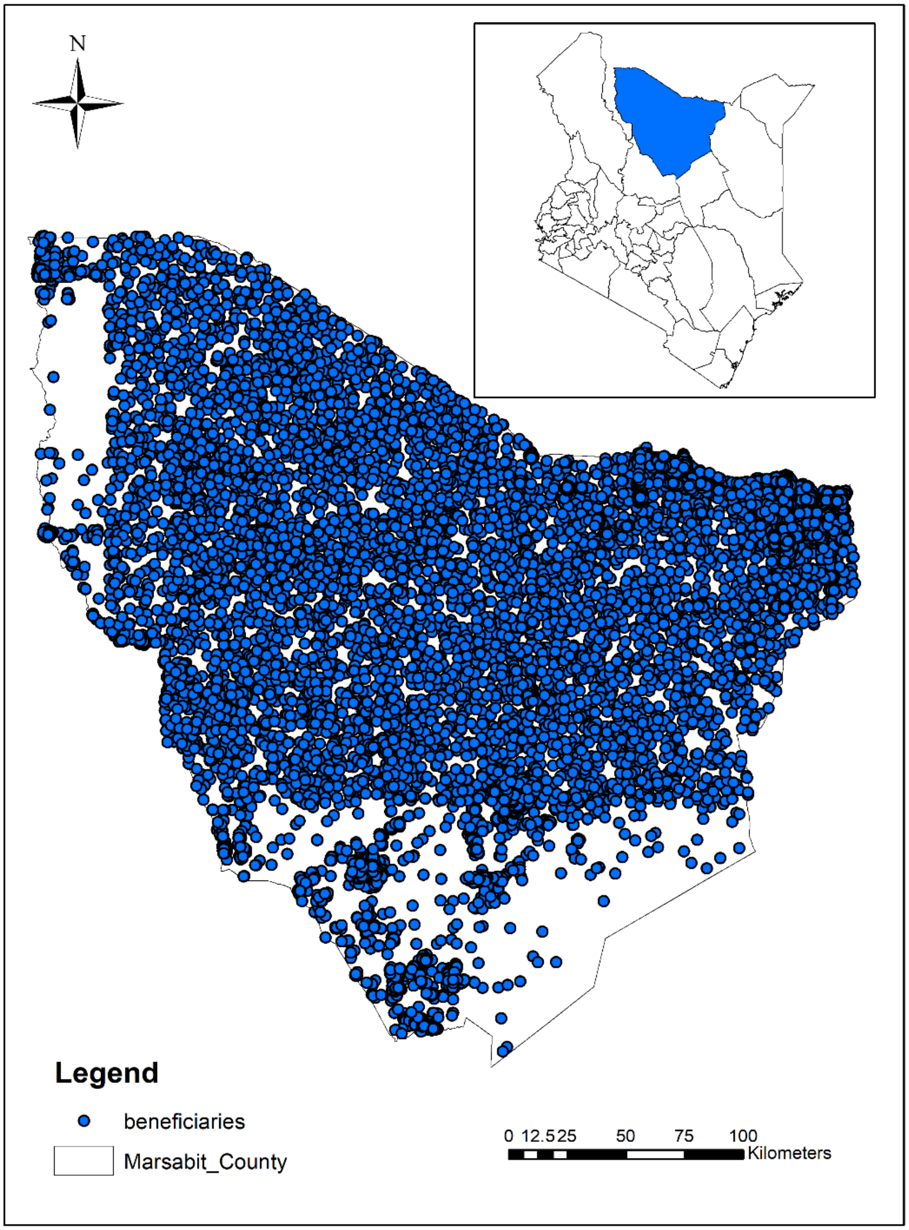
I repeat this step till I was not able to identify outliers.





Lastly, I imported the Marsabit boundary shapefile into a Geodata Frame, and defined a function to check if a coordinate is within the boundary, and iterated through the data frame to update coordinates that are not within Marsabit. It replaced those coordinates with randomly generated coordinates within the Marsabit boundary.

HSNP Marsabit Beneficiaries



For the rest of the other counties I used the above methods to clean the latitude and longitude columns.