Requested Feedback:

How much more cleaning is needed?
Feedback on how much data is missing
What kind of analysis can we do beyond basic descriptive stats and the suggested questions?

Variables chosen originally:
CountryofAsylum, nationality, ethnicity, intentionmove, TotalAdult, TotalMinor, AppliedRefugee, documentation_id, Childinschool, Childvirtualed, Childwhynotschool_nodocs, Childwhynotschool_nomoney, Childwhynotschool_failedschool, Childwhynotschool_childwork, Childwhynotschool_fearschool, Childwhynotschool_disease, Childwhynotschool_disability, Childwhynotschool_helphome, Childwhynotschool_familynot, Childwhynotschool_noschools, Childwhynotschool_nointerest, Childwhynotschool_pregnancy, Childwhynotschool_nospot, Childwhynotschool_notransport, Childwhynotschool_discrimnation, Childwhynotschool_discrimethnic, Childwhynotschool_finished, Childwhynotschool_recentlyarrive, Childwhynotschool_intransit, Childwhynotschool_noinfo, Childwhynotschool_toolate, Childwhynotschool_nolanguage, disabilityacc_disabilityserv, medicalReceived, RiskYes, RiskReturn_medical, RiskStay_medical, whynotMedical_nomoney, whynotMedical_noinsurance, whynotMedical_noID, whynotMedical_noinfo, whynotMedical_feararrested, whynotMedical_distance, whynotMedical_notavailable, whynotMedical_denied

All datasets should contain variables colored in blue. But some datasets may lack some variables colored in red

1. Created separate dataframes titled: Demo, Education, Medical
    a. Education and medical contain variables of why individuals are not going to school or getting medical care
        i. Missing data for these variables were just dropped
    b. Demo includes: CountryofAsylum, nationality, ethnicity, intentionmove, TotalAdult, TotalMinor, AppliedRefugee, documentation_id, Childinschool, Childvirtualed,disabilityacc_disabilityserv, medicalReceived, RiskYes, RiskReturn_medical, RiskStay_medical

Percent missing from Demo data across all countries and years:

Quarter	0.000000
CountryOfAsylum	0.418448
nationality	3.128073
ethnicity	65.918185

intentionmove	22.180047
TotalAdult	2.492397
TotalMinor	3.112066
AppliedRefugee	36.530766
Childinschool	55.358654
Childvirtualed	64.527931
disabilityacc_disabilityserv	6.665447
medicalReceived	54.809869
RiskYes	31.687742
RiskReturn_medical	1.294217
RiskStay_medical	6.214986

Data Cleaning Steps:

1. Removed columns with significant number of missing values and are not useful for analysis
    a. Ethnicity
    b. Disabilityacc_disabilityserv
    c. RiskStay_medical
2. Imputing other variables
    a. TotalAdult, TotalMinor
        i. Distribution is not heavily skewed, impute missing values using the median.
    b. CountryOfAsylum
        i. Certain nationalities are only found in specific CountryOfAsylum, used this relationship to impute missing values
3. Add Not Applicable for data that is "missing" because it depends on a value for another variable
    a. ChildVirtual and Childinschool
        i. For rows where TotalMinor = 0 , changed value to be Not Applicable instead of NA for clarity (ie has no value because they have no children, not because they didn't answer the question)
        ii. For rows where TotalMinor > 0, imputed values using most frequent response
4. Create Unknown for unanswered questions
    a. Intentionmove, AppliedRefugee, MedicalReceived, RiskYes
        i. Replace NA for unknown for clarity

5. Add a new column as "Quarter" to help identify the timeline

6. Merged dataset
    a. To facilitate our analysis, we initiated the process by importing datasets corresponding to identical countries across different time periods using the “read_csv” function from the pandas library in Python. We then consolidated these datasets by concatenating them and introduced a unique 'index' for each data entry to ensure traceability.
    b. In our pursuit of a dataset enriched with demographic specifics, we meticulously selected variables such as 'Year', 'Quarter', 'CountryOfAsylum', and various other identifiers that capture the essence of our demographic data. This curated subset was then preserved as a new CSV file, aptly named “HFS_country_Demo,” to signify the demographic focus within the specified country context.
    c. To address the educational and medical challenges reflected in our data, we established two distinct dataframes: "df_notschool" and "df_nomedical." These were designed to encapsulate the reasons behind the lack of school attendance and medical services access, respectively. Initially presented in a wide format in the source dataset from the UN Library, the responses were binary-coded: a "1" indicated affirmation of a reason preventing access to the required services, while a "0" signified its irrelevance.
    d. To make this data more accessible and analytically friendly, we transformed it into a long format. This was achieved by harnessing the 'melt' function from pandas, which allowed us to filter rows where the response was affirmative ('Agree' == 1) and subsequently document the specific reasons in a newly forged "Reason" column.
    e. A similar transformation was applied to the medical attention dataset, resulting in two long-format dataframes that succinctly conveyed the barriers to education and healthcare. These refined dataframes were then exported as CSV files with titles “HFS_Country_School_Reason” and “HFS_Country_Medical_Reason,” thereby completing the process of merging datasets for each country and elucidating the reasons behind the pressing educational and medical issues.

Analysis Questions

1. When family members migrate too, does it influence their intention to stay permanently in the country they migrated to? (Correlation between total number of group (total minor + total adult) with intention move)
2. What is the correlation between disability and access to healthcare? (eg. Does disability severity influence access to healthcare? (Correlation between Risk_yes and medical reason)
3. Does the intention to move predict if they are in school or not?

4. Does the decision to apply for asylum affect whether they access school/medical care?
5. Does perception of risk predict if they decide to apply for asylum or refugee status in the current country? (Risk_yes and Applied_Refugee)

Variable Description

CountryOfAsylum: Country of asylum: nation where the person or group has found safety and refuge.
nationality: This variable captures the nationality of the respondents
ethnicity: Analyze the ethnic composition of the refugee population and its relationship with various outcomes, such as access to services.
intentionmove: Explore the relationship between respondents' intention to move from their current residence and their reasons for wanting to move.
TotalAdult
Adult members in the household
TotalMinor
Number of minors in the household
AppliedRefugee Intend to apply for asylum or refugee status with current country
Documentation_id: Personal documentation do you have with you?/Valid identity card o document (ID)
Childinschool - Are the children enrolled in school?
Childvirtualed - Have resources and access to receive their education virtually
Childwhynotschool_nomoney - Reasons children not in school/Lack of financial resources
Childwhynotschool_nodocs - Reasons children not in school/Lack of documents or requirements
Childwhynotschool_noinfo
Childwhynotschool_childwork
Childwhynotschool_notransport
Childwhynotschool_disability: Reasons children not in school/Disability
Childwhynotschool_discrimethnic :ethics discrimination
Disabilityacc_disabilityserv: Disability: difficulty accessing services/Communication services for disability
medicalReceived - Feel like receiving the required medical attention?
RiskYes - At risk if they returned home
RiskReturn_medical - What risks if returning/Not being able to access medical services
RiskStay_medical - Risks if staying home/Not being able to access medical services

whynotMedical_denied
Reasons could not access required attention/I was denied medical attention
whynotMedical_distance

Reasons could not access required attention/The health center is too far
whynotMedical_feararrested
Reasons could not access required attention/Fear of being arrested or detained
whynotMedical_noID
Reasons could not access required attention/Lack of necessary documentation
whynotMedical_noinfo
Reasons could not access required attention/Lack of information or knowledge
whynotMedical_noinsurance
Reasons could not access required attention/Lack of medical insurance
whynotMedical_nomoney
Reasons could not access required attention/Lack of financial resources
whynotMedical_notavailable
Reasons could not access required attention/The medical service not available