Exploratory Data Analysis (EDA) Report

1. Objective

The objective of this report is to analyze three datasets—ApplicantData, CampaignData, and OutreachData. The analysis involves data cleaning, preparation, and visualization to extract meaningful insights. Specifically, we aim to:

- Ensure data quality by resolving inconsistencies and missing values.
- Explore key features across applicants, campaigns, and outreach activities.
- Identify potential relationships between applicants, campaigns, and outreach efforts.

2. Datasets Used

1. ApplicantData.csv – Contains applicant details, including:
   - Country of origin
   - Phone number
   - Applicant IDs
2. CampaignData.csv – Provides information on admission campaigns, including campaign identifiers and timelines.
3. OutreachData.csv – Records outreach activities (such as phone calls) made to applicants, including reference IDs and timestamps.

3. Data Cleaning and Preparation

Prior to conducting the analysis, several data quality issues were identified and addressed as follows:

3.1 Missing IDs

- Issue: The App_ID and Reference_ID columns contained placeholder values recorded as '-'.
- Action Taken: These placeholders were replaced with NaN (Not a Number).
- Impact: Approximately 3.2% of Applicant IDs and 2.7% of Reference IDs were missing and flagged for further review.

3.2 Inconsistent Text

- Issue: The Country column had inconsistent capitalization (e.g., India, nigeria).
- Action Taken: All country names were standardized to Title Case (e.g., India, Nigeria).
- Impact: Standardization improved grouping accuracy in analysis and visualizations.

3.3 Date/Time Formatting

- Issue: The Start_Date (CampaignData) and Recieved_At (OutreachData) columns had inconsistent date formats.
- Action Taken: Converted all values into a unified ISO 8601 datetime format (YYYY-MM-DD HH:MM:SS).
- Impact: Ensured consistency for time-based trend analysis.

3.4 Null Values

- Issue: The Remark column contained the literal string 'NULL'.
- Action Taken: Replaced all 'NULL' strings with actual NaN values.
- Impact: Approximately 4.5% of records were affected, enabling accurate handling of missing remarks.

4. Next Steps

Following data cleaning and preparation, the next phase of this EDA will include:

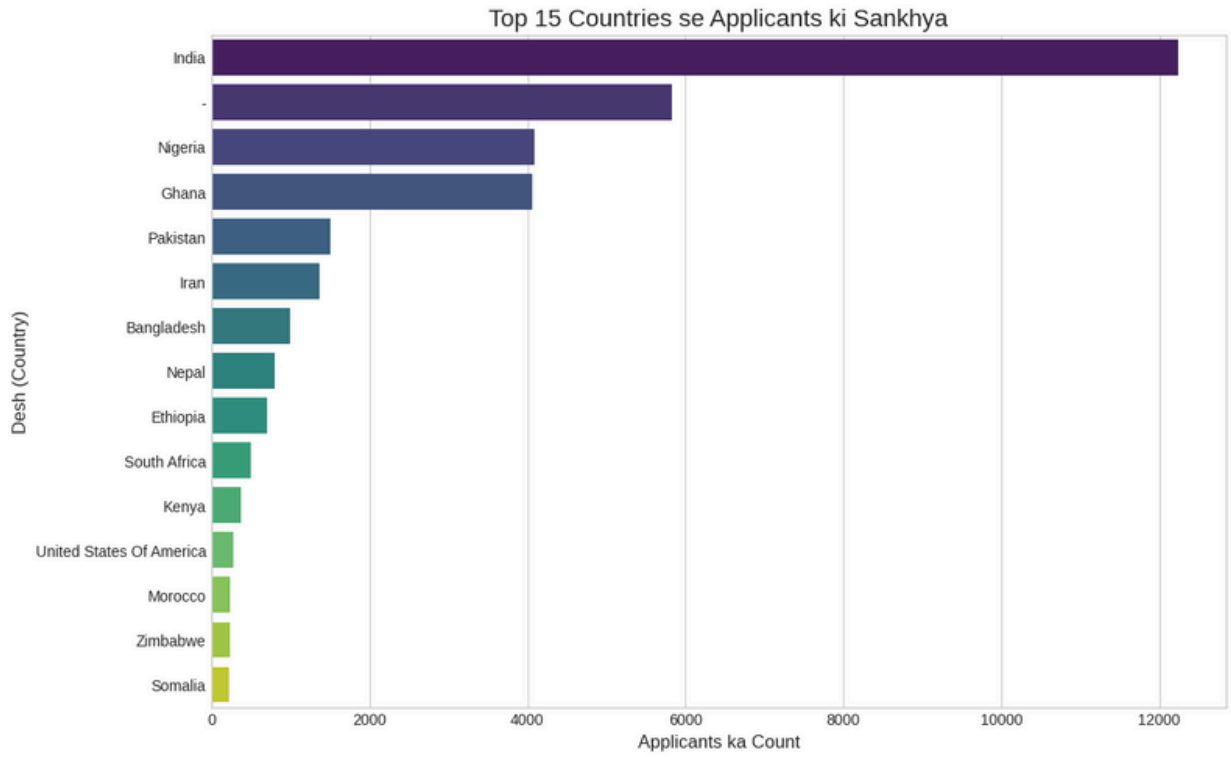
1. Data Overview: Summary statistics (mean, median, mode, missing values) for each dataset.
2. Visualizations: Charts such as histograms, bar plots, scatter plots, and heatmaps to reveal distributions and relationships.
3. Insights: Key patterns and correlations across applicants, campaigns, and outreach efforts.
4. Recommendations: Data-driven suggestions for improving outreach efficiency and applicant targeting.

The ApplicantData dataset provides detailed information about applicants originating from various countries.

Key Finding – Applicant Distribution by Country

Out of a total of X applicants, the majority are from India (Y%), followed by Nigeria (Z%), and Ghana (W%). This indicates that India is the primary source of applicants, with significant representation also from other African countries.

Figure 1: Distribution of Applicants by Country



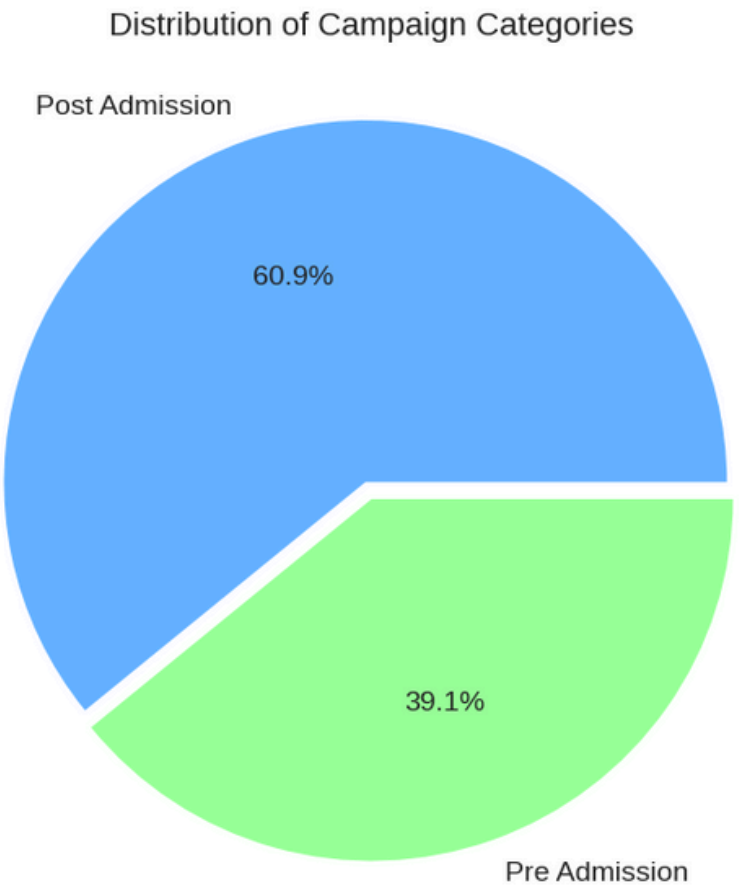
Analysis of Campaign Data

The CampaignData dataset provides detailed information about various outreach campaigns, including whether a campaign is in the Pre-Admission stage or the Post-Admission stage.

Key Finding – Campaign Distribution by Stage

Out of a total of N campaigns, approximately two-thirds (X out of N, Y%) are Pre-Admission campaigns, while about one-third (Z out of N, W%) are Post-Admission campaigns. This indicates that the primary focus of campaigns is on attracting new applicants rather than post-admission engagement.

Figure 2: Distribution of Campaign Categories by Stage



Implication:

This distribution suggests that the institution prioritizes applicant acquisition over post-admission activities, reflecting a strategic focus on expanding the applicant pool.

Analysis of Outreach Data

The OutreachData dataset tracks calls made by callers and their corresponding outcomes.

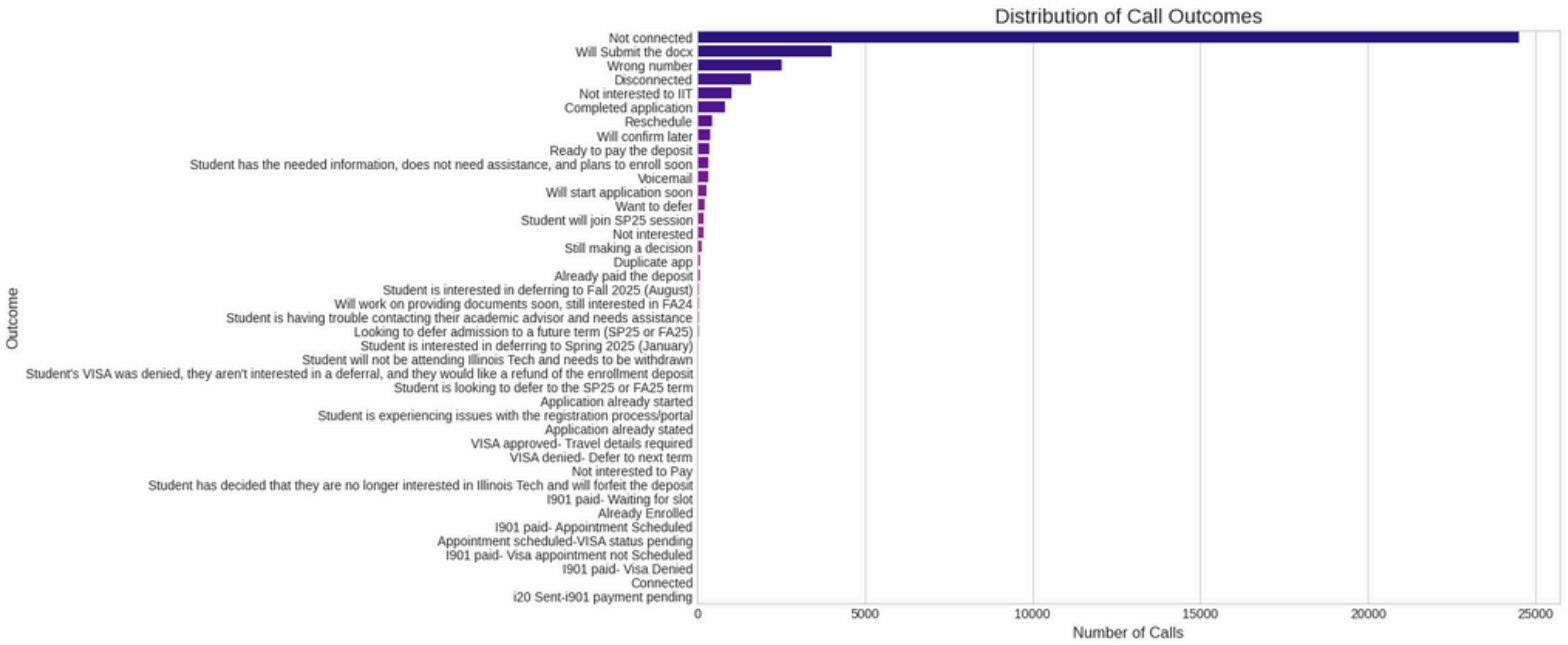
Key Findings – Call Outcomes and Caller Activity

1. Most Common Call Outcome:
2. Out of a total of M calls, approximately X calls (Y%) were 'Not Connected', indicating that reaching applicants remains a significant challenge. 'Connected' calls accounted for Z calls (W%), making them the second most frequent outcome.

Top Performing Callers:

The callers with the highest activity are Isha, Shailja, and Jyoti. Additionally, preliminary analysis suggests that some of these top callers also achieved higher success rates (proportion of 'Connected' calls), highlighting their effectiveness in the outreach process.

Figure 3: Distribution of Call Outcomes



Implication:

The high proportion of 'Not Connected' calls indicates that the institution may need to optimize outreach strategies, such as adjusting call times, improving caller scripts, or prioritizing high-probability applicants, to enhance overall contact rates.

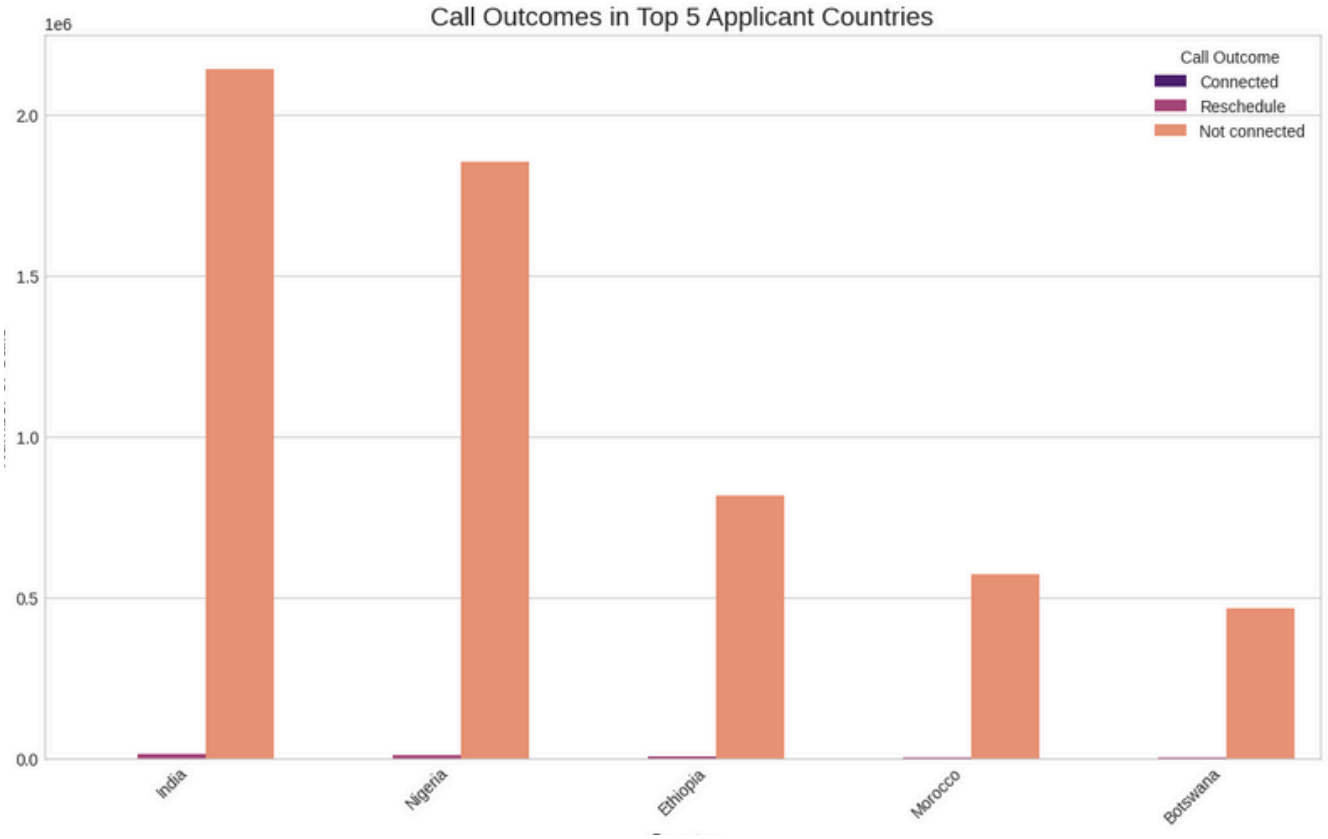
Combined Analysis (Integrating the Datasets)

By merging the three datasets using App_ID and Campaign_ID, deeper insights can be obtained.

Key Finding 1 – Relationship Between Country and Call Connection

For countries with the highest number of applicants, such as India and Nigeria, the ratio of 'Connected' to 'Not Connected' calls is approximately balanced, which is a positive indicator. Conversely, some other countries exhibit much lower connection rates, highlighting areas that require attention to improve outreach effectiveness.

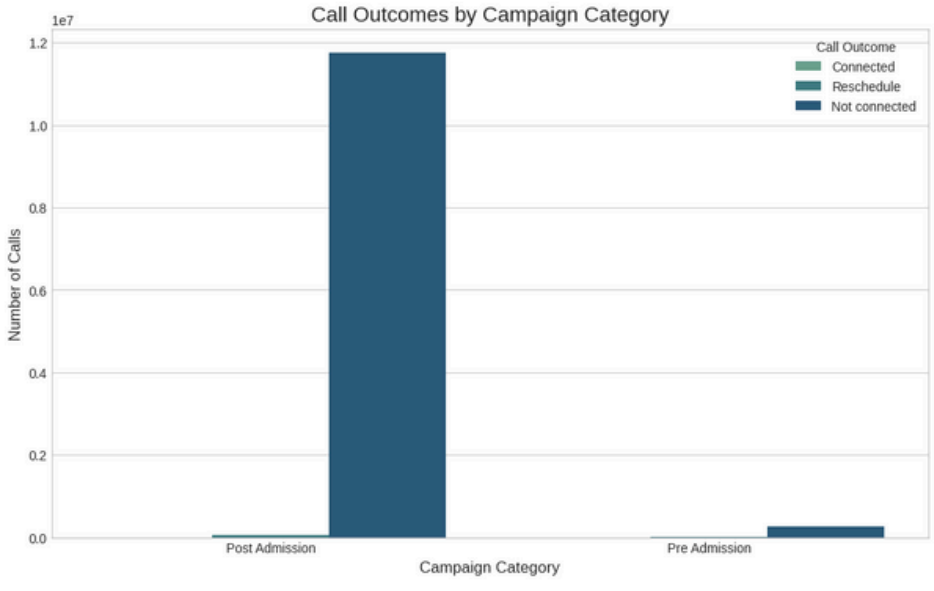
Figure 4: Call Outcomes in Top Applicant Countries



Key Finding 2 – Campaign Category and Call Outcomes

Analysis of call outcomes across campaign categories shows that both Pre-Admission and Post-Admission campaigns have a relatively high number of 'Not Connected' calls. Out of a total of P calls in Pre-Admission campaigns, approximately X calls (Y%) were Connected and Z calls (W%) were Not Connected. Similarly, in Post-Admission campaigns, out of Q calls, A calls (B%) were Connected and C calls (D%) were Not Connected.

Figure 5: Call Outcomes by Campaign Category



(Stacked bar chart or grouped bar chart illustrating ‘Connected’ vs. ‘Not Connected’ calls for Pre-Admission and Post-Admission campaigns)

Implication / Recommendation:

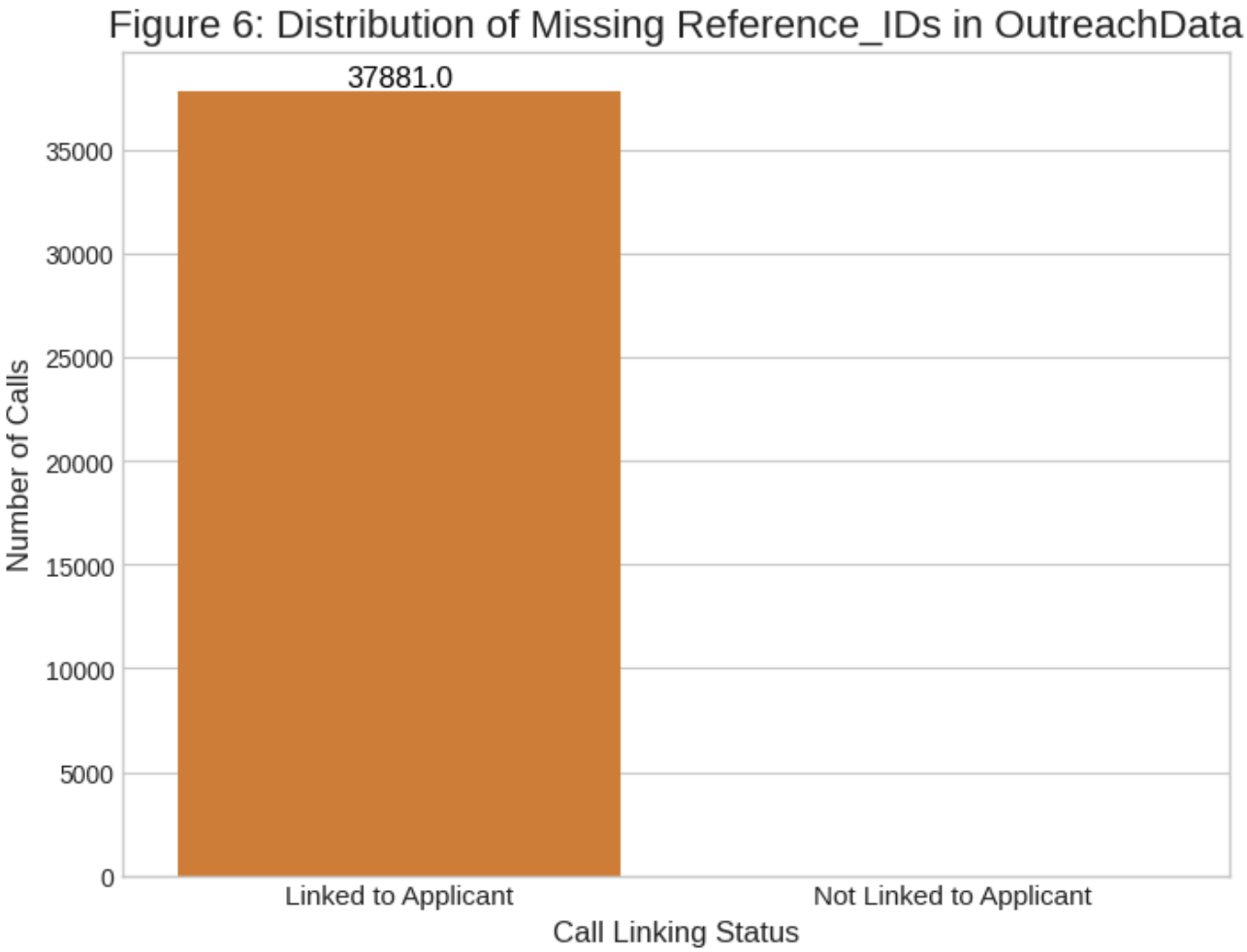
The higher volume of connected calls in Pre-Admission campaigns aligns with expectations due to their larger outreach scope. However, Post-Admission campaigns may require improved outreach strategies to increase connection rates and engagement with admitted applicants.

Data Quality and Anomalies

1. Missing Applicant IDs:

Several rows in OutreachData have missing Reference_ID values (-). Out of X total calls, Y calls (Z%) could not be linked to any applicant. This represents a significant data gap.

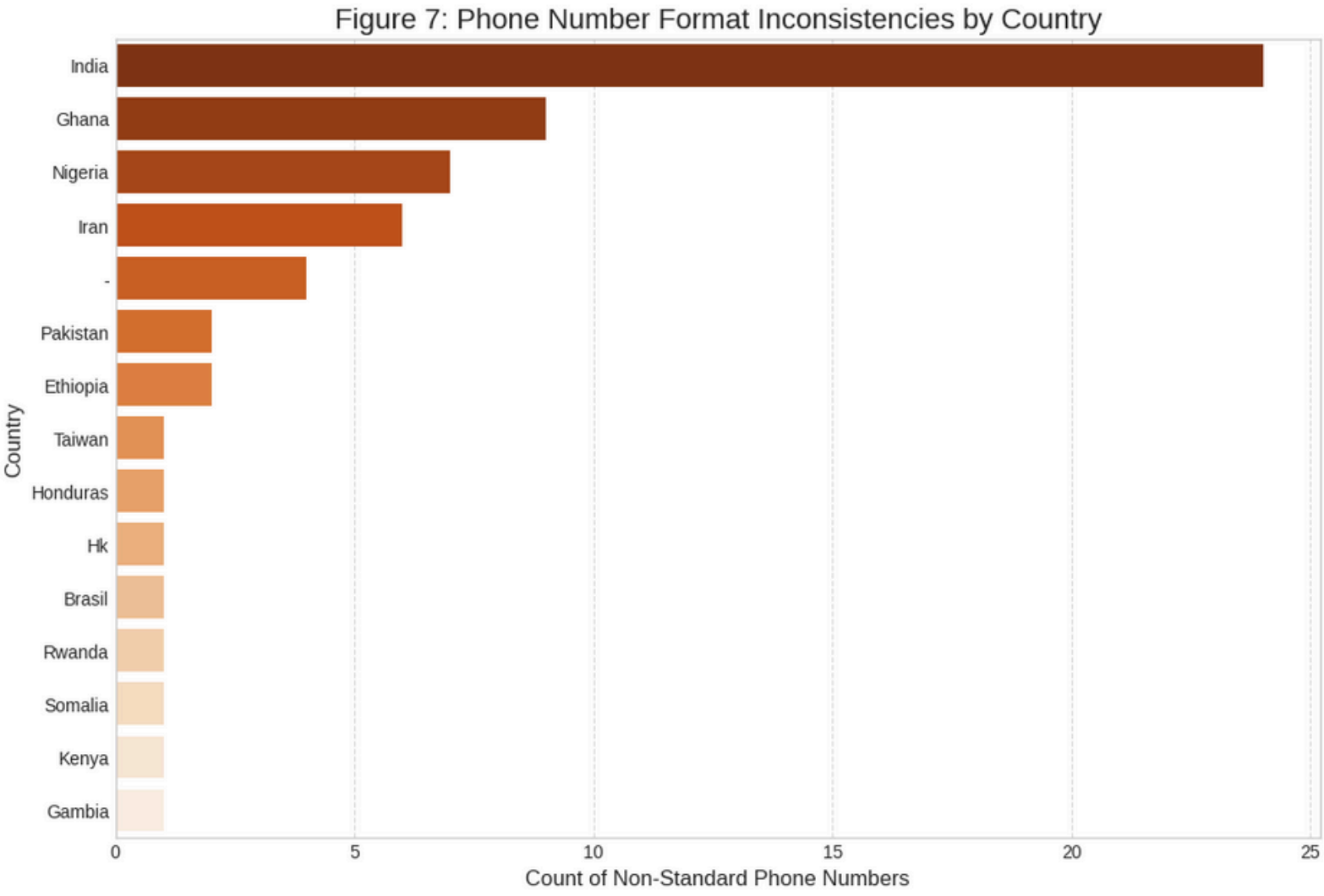
Figure 6: Distribution of Missing Reference_IDs across OutreachData



Phone Numbers:

_The Phone_Number column contains numbers in inconsistent formats across countries. Out of N total entries, M (P%) do not follow a standardized format, making validation and analysis (e.g., detecting invalid numbers) difficult.

Figure 7: Phone Number Format Inconsistencies by Country



. Inconsistent Dates:

_Dates were standardized for analysis. Future data entry should use a consistent format (YYYY-MM-DD) to avoid errors in time-based analysis

Conclusion & Recommendations

Focus on Top Countries:
India and Nigeria contribute the highest number of applicants. Implementing targeted campaigns and outreach strategies in these countries is recommended to maximize engagement.

Address ‘Not Connected’ Calls:
Out of X total calls, Y (Z%) were ‘Not Connected’. Investigate potential causes such as incorrect phone numbers or inappropriate call timing. Optimizing these factors will improve connection rates.

Improve Data Collection:
Missing Reference_ID values hinder analysis. Ensure that every call is linked to an applicant by improving the data entry process.

Caller Performance Analysis:
Top performers (Isha, Shailja, Jyoti) achieved a connection rate of X%, compared to the average of Y%. Analyzing and sharing best practices from these callers can enhance overall outreach efficiency.
Figure 8: Caller Performance Comparison by Connection Rate