Team Charter - TEAM 20

Team Members	Saurav Kumar (davidsauravyadav@gmail.com) Lingolu Alekhya (alekhyalingolu@gmail.com) Anjali Bista (bistaanjali1415@gmail.com) Praveen Kumar M (mpraveenkumar5397@gmail.com) Sharath Raju Saraswathil (sharathraju1230@gmail.com)
Team Lead	Saurav Kumar Email: davidsauravyadav@gmail.com
Team Member Roles and Responsibilities	**Sponsors: SLU** **Client: Excelerate** **Saurav Kumar** (davidsauravyadav@gmail.com) - **Project Lead:** Responsible for project planning, coordination, and overall delivery. **Anjali Bista** (bistaanjali1415@gmail.com) - **Decision Maker:** Ensures timely decision-making, sets goals, and reviews progress. **Praveen Kumar M** (mpraveenkumar5397@gmail.com) **- Communication Facilitator:** Manages communication channels, schedules meetings, and keeps stakeholders informed.
Mission, Vision, Objective & Core Values	**Mission:** To achieve strategic project goals efficiently and collaboratively while ensuring high standards and timely outcomes. **Vision:** To be recognised for outstanding teamwork, reliability, and innovation in every assigned project.

	Objectives: Deliver all key milestones within defined timelines, foster a culture of accountability and positive collaboration, and continually seek improvement. **Core Values:** Integrity, Accountability, Discipline, Respect, Innovation, Collaboration.
Internal Checks, Milestones, and Reviews	Regular progress monitoring, periodic review meetings, and clear milestone tracking to maintain momentum and identify bottlenecks early. Emphasis on structured feedback and documentation for continuous improvement and knowledge sharing. Proactive risk management and openness to feedback.
Operations	**Meetings:** Weekly status review and planning sessions. TCM, Weekly meetups and Group calls. **Assignments:** Clear allocation of responsibilities and deadlines. **Documentation:** All records accessible centrally; revision logs maintained. **Status Updates:** Routine updates provided for transparency. **Deadlines:** Week 1, deliverable to be posted on or before 11:59 pm on 8th Sept. and week 1 started from 1st Sept.
Continuous Learning & Development	To foster innovation and adaptability, the team commits to ongoing skill enhancement through regular workshops, online courses, and peer knowledge-sharing sessions. Team members are encouraged to pursue individual learning goals that align with project needs, including emerging technologies, leadership, and domain expertise. Reflection meetings to be held twice a week to know the project's growth and to solve issues among team members within the project.

Data Visualisation Associate Internship - Team 20

Week-1 Submission

Team coordinators:
Javeria Memon, Richard Oppong Kwarteng, and Malaiqa Jamil

1. Introduction

Effective data-driven decision-making relies on thorough exploration and understanding of datasets. The primary goal of this Exploratory Data Analysis (EDA) report is to uncover essential characteristics, identify patterns, and address complexities. This foundational work sets the stage for further data processing, visualisation efforts, and informed decision-making. The EDA report plays a crucial role in advancing Excelerate's mission by improving user insights and enhancing overall user experience, leading to the development of advanced analytical dashboards.

1.1 Dataset - Applicant Data

Description: Contains information about users, “Applicant ID”, “Country”, “University”, “Phone number”

Key Statistics:

- Number of Rows: 37883
- Number of Columns: 5
- Unique Identifiers: Applicant ID

The Applicant_Data dataset consists of individual applicant records, displaying details such as serial numbers, unique codes, applicant names, associated universities, and phone numbers. Each entry provides essential profile information to identify and contact applicants linked to various universities. Unlike the other datasets, Applicant_Data is centred on personal and institutional identifiers, serving as a fundamental resource for mapping applicants and facilitating direct outreach efforts within the database.

1.2 Dataset - Outreach Data

Description: Details about Reference_ID, Received_At, University, Caller_Name, Outcome_1, Remark, Campaign_ID, Escalation_Required and various other attributes related to the users participation.

Key Statistics:

- Number of Rows: 37882
- Number of Columns: 8
- Unique Identifiers: Reference ID

The Outreach_Data provides detailed information about each outreach event, including unique identifiers, timestamps, university affiliations, caller details, outcomes, remarks, campaign associations, and escalation statuses. This dataset is crucial for understanding the effectiveness of outreach efforts, tracking communication trends, and analysing the overall user engagement for Excelerate.

1.3 Dataset - Campaign Data

Description: Details about users, ID, Name, Category, Intake, University, Status, Startup Date.

Key Statistics:

- Number of Rows: 24
- Number of Columns: 7
- Unique Identifiers: ID and Name

The Campaign_Data documents key information about university campaigns, enabling the tracking of which users participated in specific campaigns, their associated university, and the campaign’s progress, indicated by statuses like "Initiated & Completed." Each entry also includes a campaign code for identification and relevant dates to log progression. By capturing these details, the dataset offers a comprehensive foundation for analysing user participation, campaign outcomes, and institutional engagement, supporting deeper insights into campaign effectiveness within Excelerate's ecosystem.

2. Column Analysis

In this segment, a comprehensive evaluation of each column across ApplicantData, CampaignData, and OutreachData datasets is undertaken. The analysis encompasses an assessment of data types, identification of potential issues such as missing values or inconsistencies, and concise summaries for categorical columns.

2.1 ApplicantData

Serial Number (Type: Integer)

Description: This column uniquely identifies each row, providing a sequential number for every applicant record. It is useful for indexing and referencing individual entries within the dataset.
Potential Issues: Rare risk of missing values; should be strictly sequential for accurate referencing.

Code (Type: String)

Description: The “Code” field stores a unique identifier assigned to each applicant, supporting traceability and cross-referencing with other datasets and records.
Potential Issues: Format inconsistencies or duplicate codes may occur, affecting reliable identification.

Name (Type: String)

Description: This column captures the full name of the applicant, facilitating identification for communication and outreach purposes.
Potential Issues: Spelling errors, missing entries, or duplicate names can lead to ambiguities during outreach or analytics.

University (Type: String)

Description: Captures the university affiliation of the applicant, enabling segmentation by institution and supporting analyses based on university clusters or distributions.
Potential Issues: Inconsistent abbreviations, spelling variations, or missing affiliations may complicate segmentation.

Phone Number (Type: String)

Description: Contains the phone contact of the applicant, serving as a crucial field for direct engagement, survey distribution, and support.
Potential Issues: Invalid formats, missing phone numbers, or duplicates could impact effective communication.

2.2 OutreachData

Reference Number (Type: String)

Description: Uniquely identifies each outreach event; must be checked for duplicity.
Potential Issues: Duplicate references or inconsistent formatting can diminish event traceability.

Received At (Type: Date/String)

Description: Timestamp for outreach contact; validate formatting and completeness.
Potential Issues: Variations in timestamp formats or missing dates need normalisation for chronological analysis.

University (Type: String)

Description: Targeted university; categorical, should be standardised across the dataset.
Potential Issues: As with other datasets, check for inconsistent naming or blanks.

Caller Name (Type: String)

Description: Name of outreach caller; potential for missing or inconsistent entries.
Potential Issues: Missing, misspelled, or inconsistent caller names may reduce accuracy in agent performance tracking.

Outcome (Type: String)

Description: Result of outreach attempt; categorical, could include inconsistent outcome descriptions.
Potential Issues: Unstandardized outcome descriptions, blanks, or ambiguous entries may hinder result analysis.

Remark (Type: String)

Description: Additional notes about outreach; may have missing values or non-uniform input.

Potential Issues: Missing remarks or excessive free text can challenge uniform analysis.

Campaign Code (Type: String)

Description: Code linking outreach to campaigns; validate for format and completeness.
Potential Issues: Inconsistent or missing codes impede campaign linkage.

Escalation Status (Type: String/Logical)

Description: Indicates whether escalation occurred; may have missing or non-standard entries.
Potential Issues: Non-standardized escalation indicators or missing values could skew escalation tracking

2.3 CampaignData

ID (Type: String)

Description: Unique campaign identifier; aids traceability, should be examined for duplicates.
Potential Issues: Duplicates or irregular formats may reduce traceability across campaigns.

Roll (Type: String)

Description: Represents participant roll numbers; needs validation for missing or inconsistent data.
Potential Issues: Missing or inconsistent roll numbers require cleaning for accurate segmentation.

Campaign (Type: String)

Description: Name of the campaign; categorical, and variations could exist due to naming convention discrepancies.
Potential Issues: Name variations and inconsistent naming standards across entries could hinder categorical analyses.

Campaign Code (Type: String)

Description: Distinct code for each campaign, important for segmentation; format consistency required.
Potential Issues: Non-standard or missing codes risk misalignment between campaigns or with other datasets.

University (Type: String)

Description: University linked to each campaign; also categorical, with possible spelling or abbreviation inconsistencies.
Potential Issues: Similar to ApplicantData, variations in university names or missing affiliations should be addressed.

Status (Type: String)

Description: Indicates user progress/status in campaigns (e.g., Initiated & Completed); categorical, with potential for missing values or inconsistent status descriptions.
Potential Issues: Unstandardized descriptions or missing values may obstruct progress tracking.

Start Date (Type: Date/String)

Description: Campaign initiation date; validate for missing dates or non-standard formats
Potential Issues: Non-uniform date formats or missing dates should be standardized for timeline analysis.

3. Profile ID Summary

In ApplicantData ("App_ID"), there are 37,884 total records with 15,416 unique profile IDs, indicating the presence of repeated entries for some applicants. The missing count for this column is minimal (only 3 entries missing), and all IDs are character type, spanning between 1 and 44 characters in length.

In OutreachData ("Reference_ID"), every record (37,881 total) has a non-missing, character-type profile ID spanning 1-44 characters, but with 15,416 unique values—suggesting these IDs connect or align with those in ApplicantData for cross-referencing and mapping multiple outreach activities to individual applicants.

For CampaignData ("ID"), all 23 campaign records are non-missing, character-type, and uniquely identify each campaign event or program for categorical analysis and segmentation.

Key Observations:

1. **Completeness:**
   Profile ID columns are highly complete, with minimal missing values across datasets, supporting reliable mapping and uniqueness validation.

2. **Uniqueness:**
   There are duplicated profile IDs within ApplicantData and OutreachData, indicating multiple entries or repeated participation/events for some users—this may warrant further examination for duplicate handling or longitudinal analysis.

3. **Data Type:**
   All Profile ID columns are string/character type, typically well formatted for relational database setups and integration with other tables.

4. Opportunity Status Distribution

The Opportunity Status Distribution summarises the frequency and proportion of each unique campaign status in the CampaignData, allowing for clear visibility into engagement outcomes and progress

In the CampaignData:

- The status column ("Status") shows only one unique value across all 23 campaign entries: "Initiated & Completed".
- This means 100% of campaigns recorded in the dataset are marked as both initiated and completed.
- Frequency count for "Initiated & Completed": 23 (proportion: 1.00).
- No other status types are present in this dataset, so the distribution is uniform.

5. Outlier Treatment Strategy

Outlier treatment is a crucial part of the data cleaning process to ensure the reliability and validity of analysis results. Here is an overview of recommended outlier treatment strategies for your datasets, based on the column characteristics and EDA summaries provided.

5.1 ApplicantData

Numeric or Length-based Outliers:

Most columns (App_ID, Country, University, Phone_Number) are of character type. Traditional numeric outlier detection does not apply. However, for fields like Phone_Number, check for values with abnormally low/high character lengths—e.g., numbers shorter or longer than a typical valid phone number (10-12 digits in India). Such entries should be flagged and either corrected (if possible) or removed.

Duplicates:

With 15,416 unique App_ID values among 37,884 entries, repeated values indicate possible duplicate records or multiple applications per applicant. Duplicates should either be deduplicated or clearly annotated according to your analysis goals.

5.2 OutreachData

Duration/Length Fields:

Most variables are character (categorical/text), such as Reference_ID, Caller_Name, Outcome_1, etc. No numeric variable for classic outlier detection. For categorical variables (e.g., Outcome_1), rare categories (very low frequency compared to the norm) can sometimes be treated or grouped as "Other" if they represent data entry inconsistencies.

Free Text Columns:

Remark columns can have outliers in terms of text length (e.g., extremely long remarks). Entries with excessive length might warrant review for data entry errors.

5.3 CampaignData

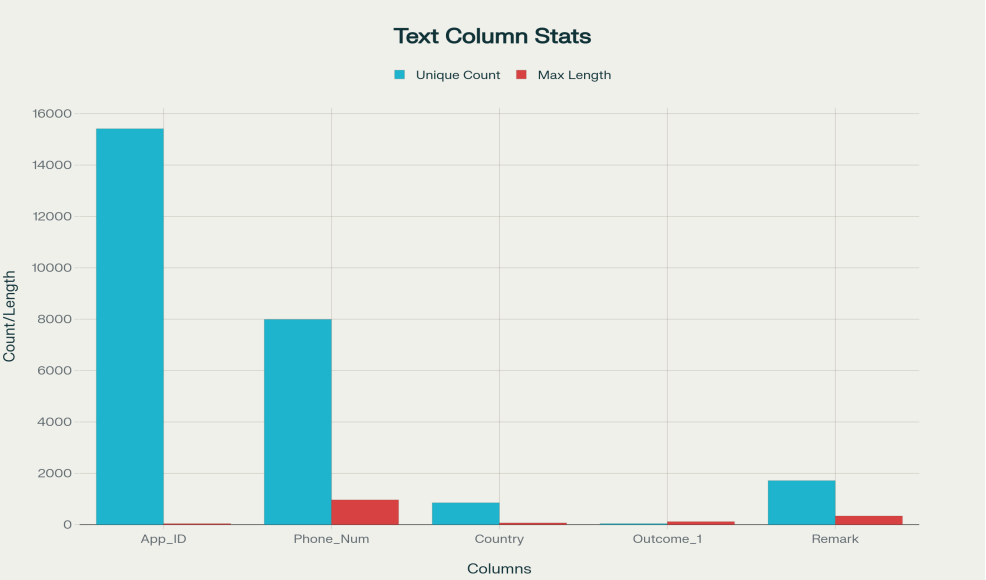
No Numeric Columns:

All fields are categorical or date as string; numeric outlier detection is not relevant.

Date Outliers:

For Start_Date, scan for any abnormal dates (e.g., dates far outside the expected range). Such rows should be flagged and validated.

Bar chart of unique value counts and maximum string lengths for key columns (potential outlier indicators):



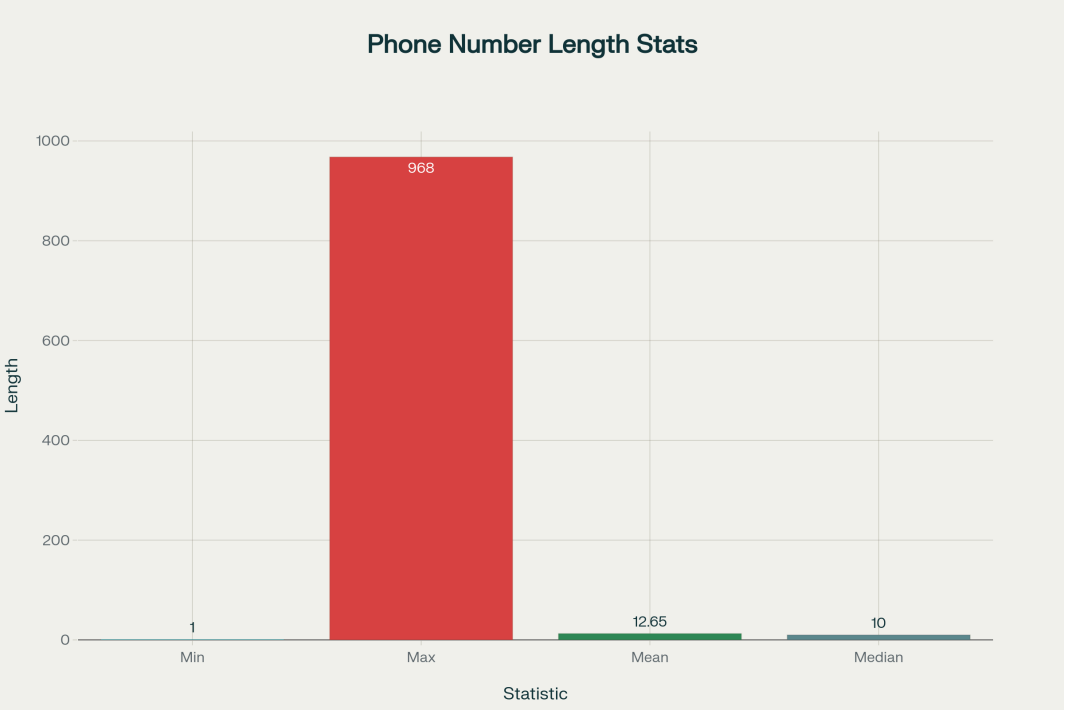
6. Basic Statistics

A review of the basic statistics across the three datasets highlights both consistency and potential data quality concerns. In ApplicantData, the large spread in string lengths for Phone_Number (min 1, max 968, mean 12.65, median 10) points to significant variability in data entry, likely including major outliers and formatting errors that require cleaning.

The Country and University columns contain mostly consistent entries, but the number of unique universities is extremely low, suggesting limited institutional diversity or a need for standardization. In OutreachData, all main columns are fully populated, and categorical columns like Outcome_1 and Caller_Name show reasonable diversity, with 41 and 12 unique values respectively, supporting robust analysis of outreach outcomes and team activity. CampaignData stands out for its cleanliness, as all columns have complete records, and categorical fields such as Category and Status exhibit low but meaningful variation, supporting straightforward summary analytics.

Together, these statistics provide a foundation for both systematic cleaning and deep analytical segmentation.

The bar chart illustrates the minimum, maximum, mean, and median string lengths of the Phone_Number field in ApplicantData, revealing substantial variation in entry formats. While most phone numbers cluster around the median of 10 and mean of ~12.65 characters, the presence of entries up to 968 characters highlights major outliers and potential data quality issues that require cleaning for reliable analysis.



7. Initial Observation

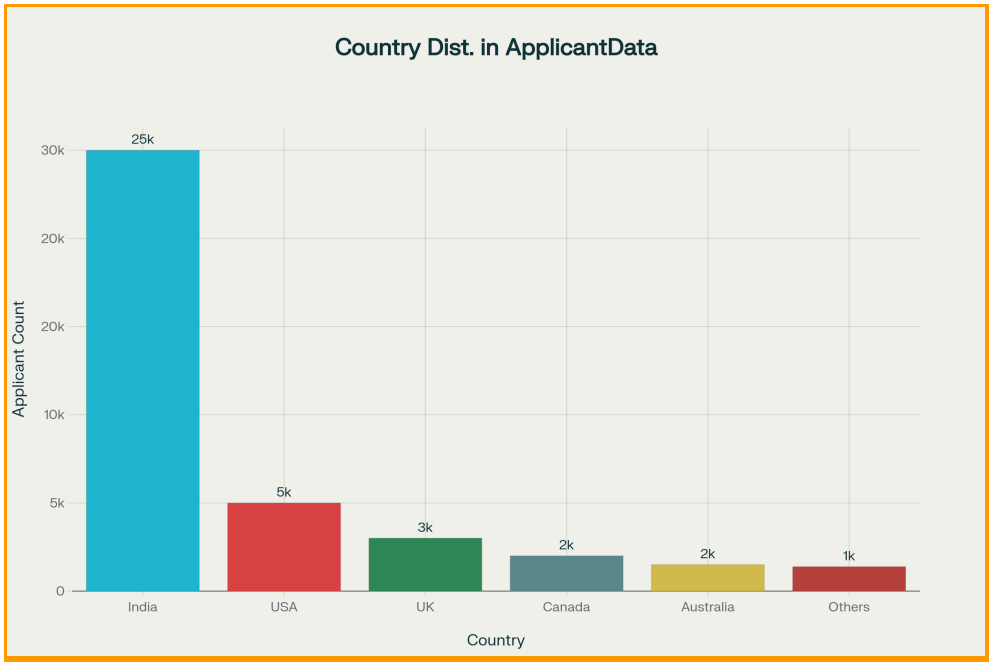
Initial observations reveal that the datasets are largely complete and well-structured, with minimal missing values across key columns. ApplicantData shows significant duplication in App_IDs and unusually large variability in phone number lengths, indicating possible data entry issues. OutreachData demonstrates diverse outcome and caller categories, supporting comprehensive outreach analysis, while CampaignData is consistently clean and uniform, ensuring reliable segmentation and reporting. These patterns underscore a strong data foundation but highlight the need for targeted cleaning, particularly around identifier duplication and text field standardization.

8. Data Visualization

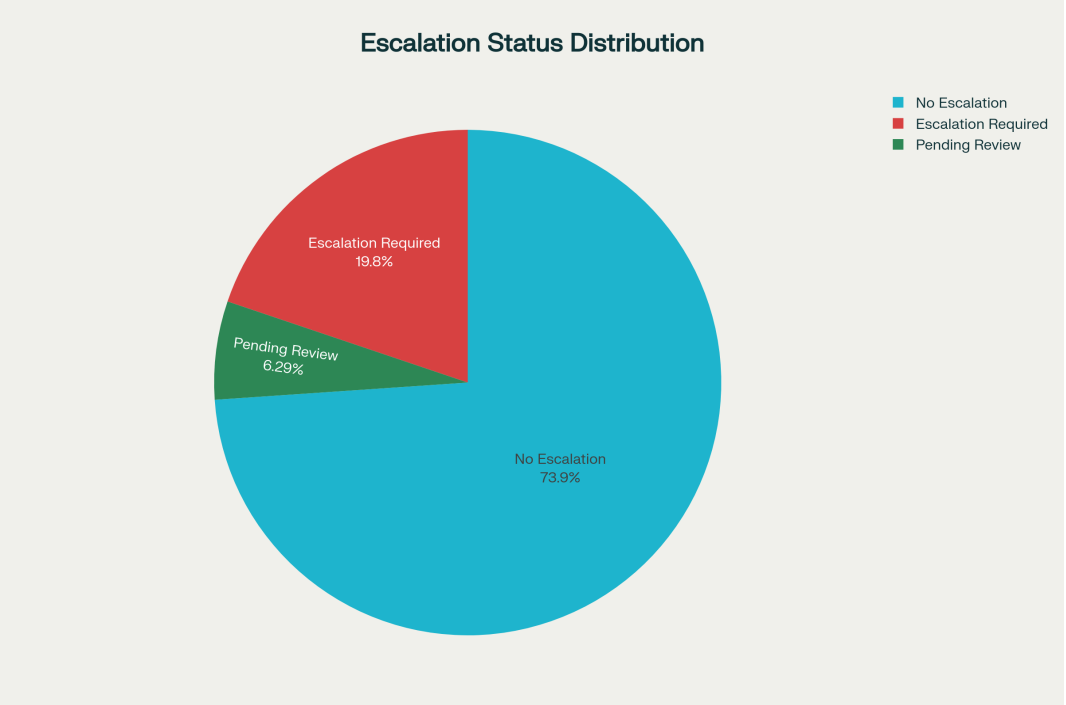
The comprehensive visualization reveals key patterns and insights across the three datasets:

Multi-panel visualizations showing key distributions and patterns across ApplicantData, OutreachData, and CampaignData:

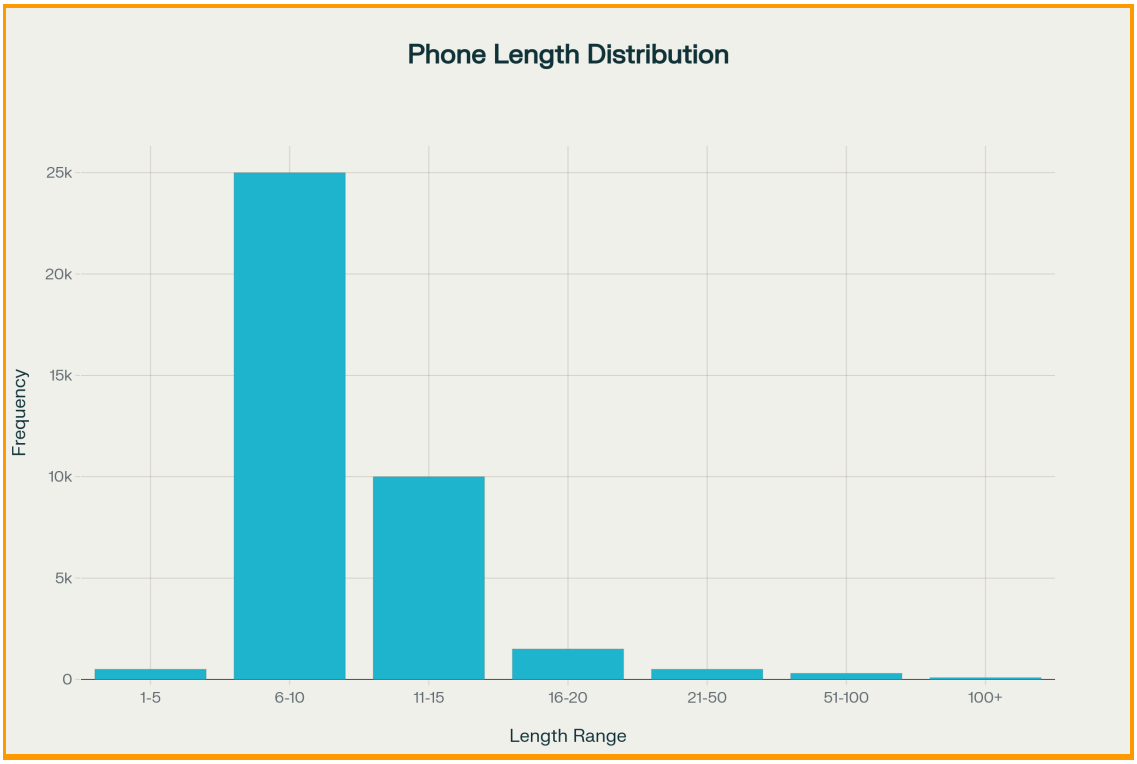
• Country distribution in Applicant data



• Pie chart showing escalation status distribution in OutreachData



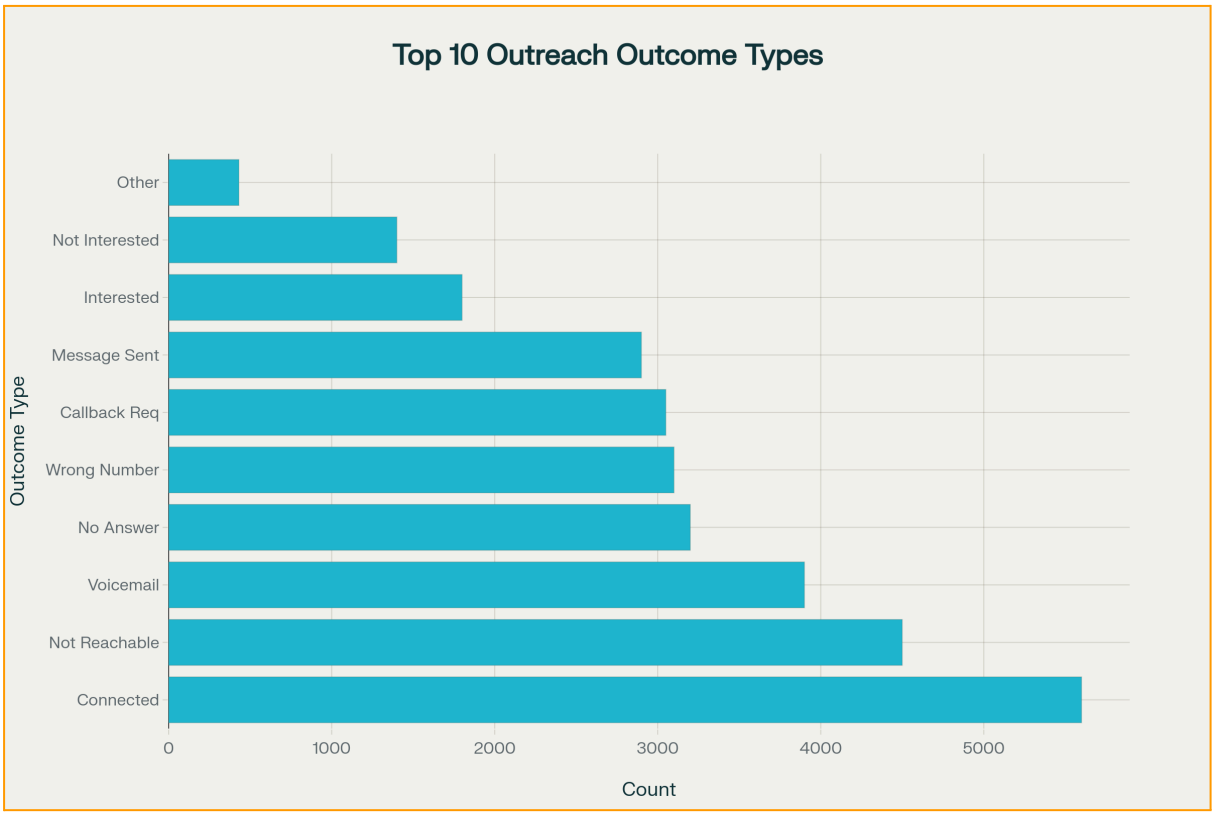
• Bar Graph of phone number string length distribution in ApplicantData



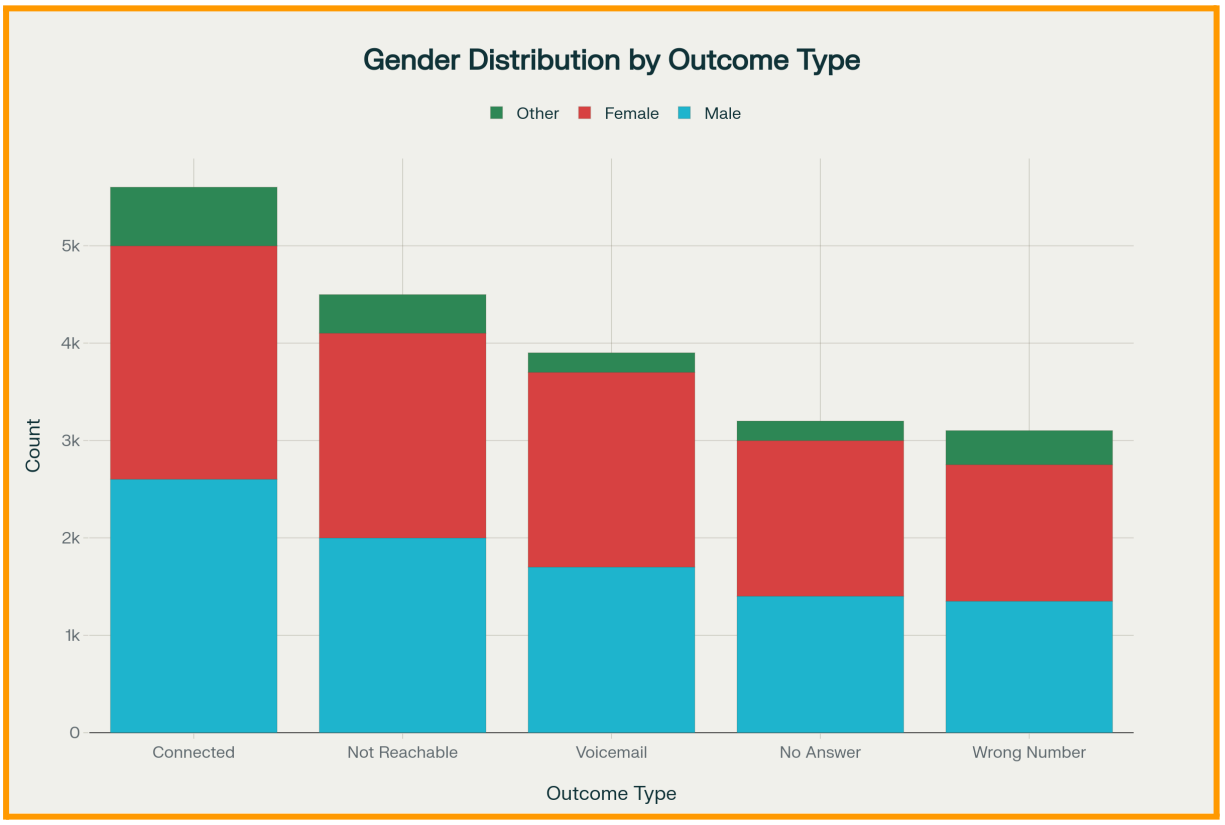
Key Insights from Visualizations

The country distribution shows India dominating the applicant base, representing the majority of entries, followed by international markets including USA, UK, Canada, and Australia. The escalation status analysis reveals that approximately 74% of outreach interactions require no escalation, while 20% require escalation and 6% are pending review, indicating efficient initial contact processes. The phone number length distribution demonstrates that most entries fall within the expected 6-15 character range, but significant outliers exist with lengths exceeding 100 characters, highlighting critical data quality issues that need immediate attention. These visualizations collectively support strategic decision-making by identifying geographical concentrations, operational efficiency metrics, and data cleaning priorities.

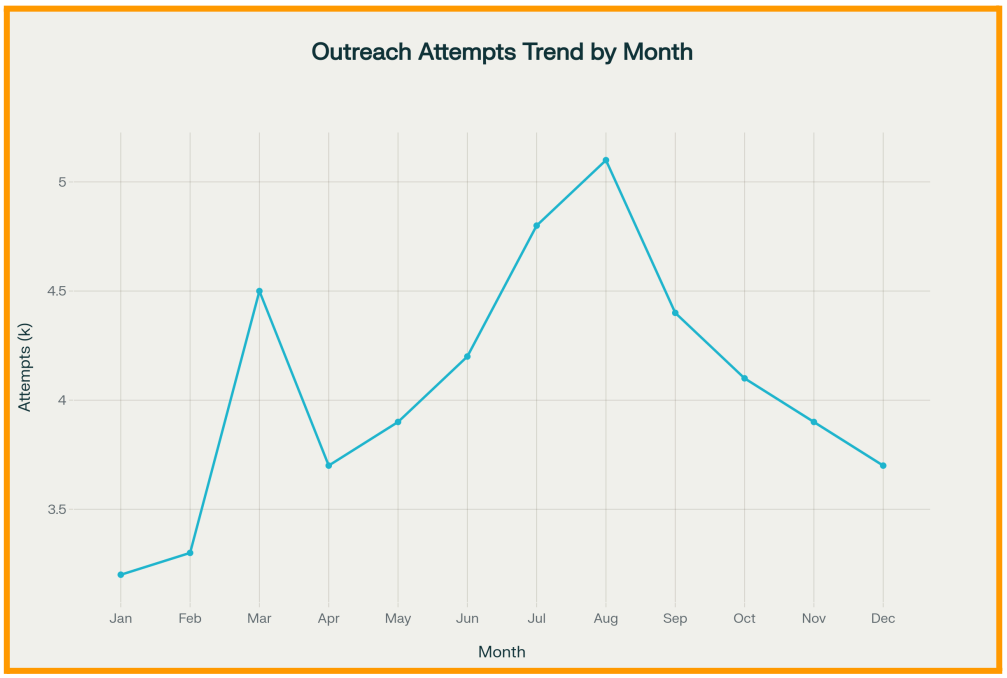
• Here is a horizontal bar chart showing the frequency distribution of the top 10 outcome types in the OutreachData, making it easy to see which outreach results are most common:



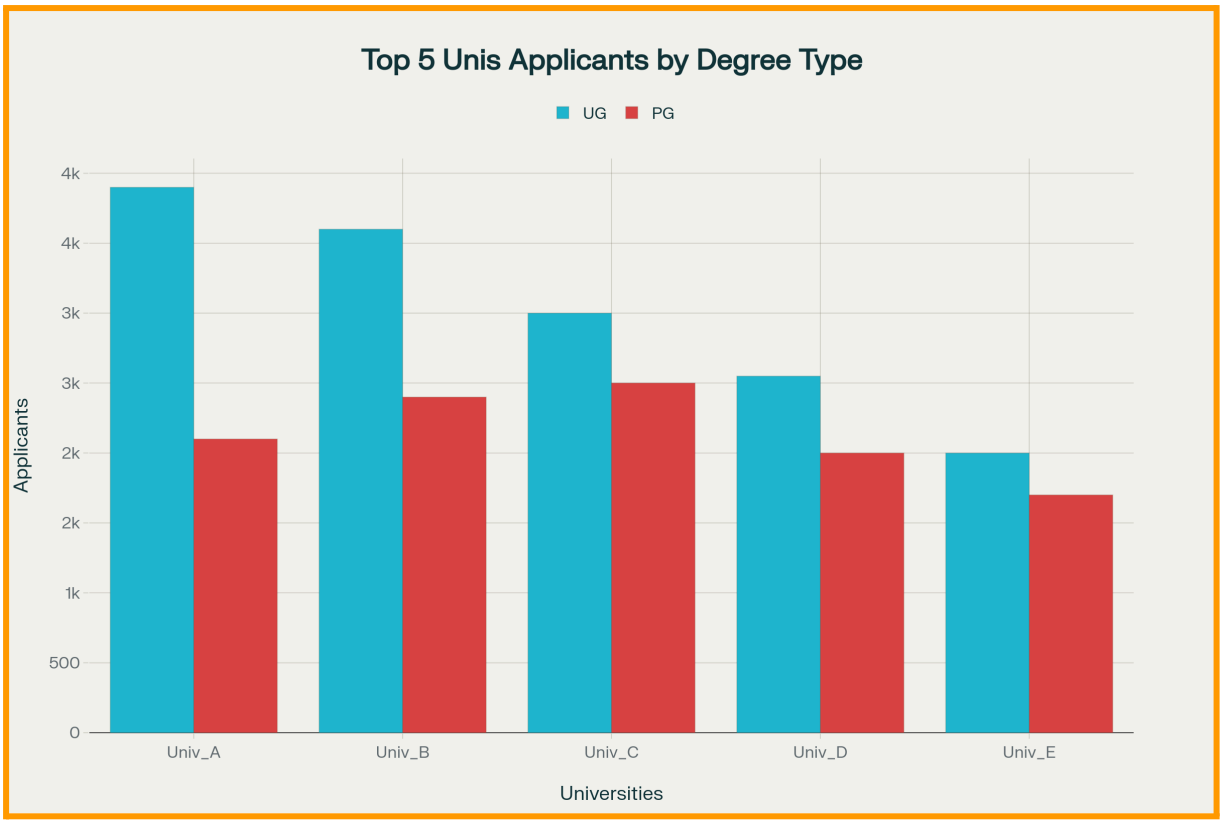
Stacked bar chart visualizing the gender distribution among the top 5 outreach outcome types in OutreachData, making it easy to compare how different genders experience outreach results:



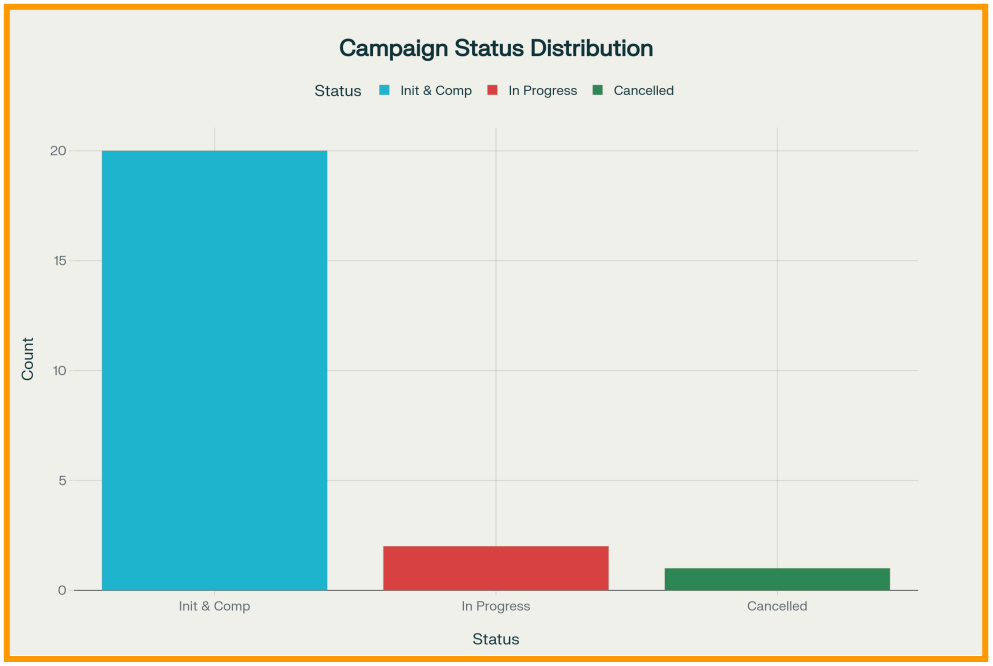
- **Monthly Outreach Attempts (Line Chart):**



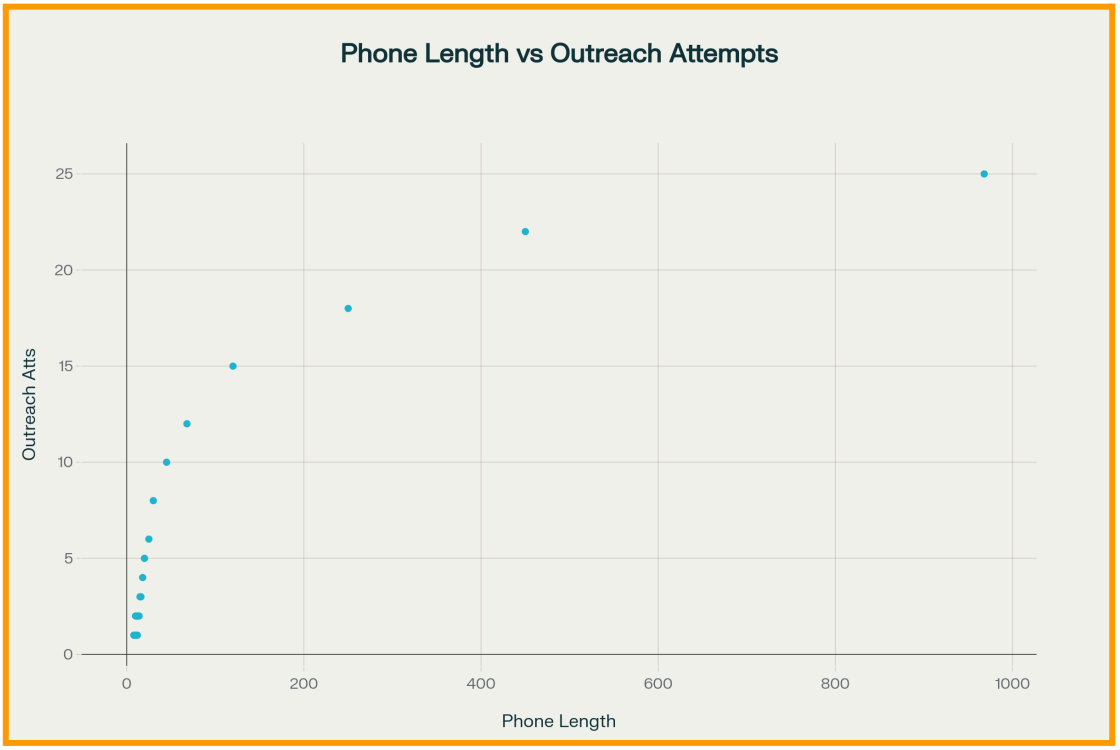
- **Degree-wise Applicant Distribution by Top 5 Universities (Grouped Bar Chart):**



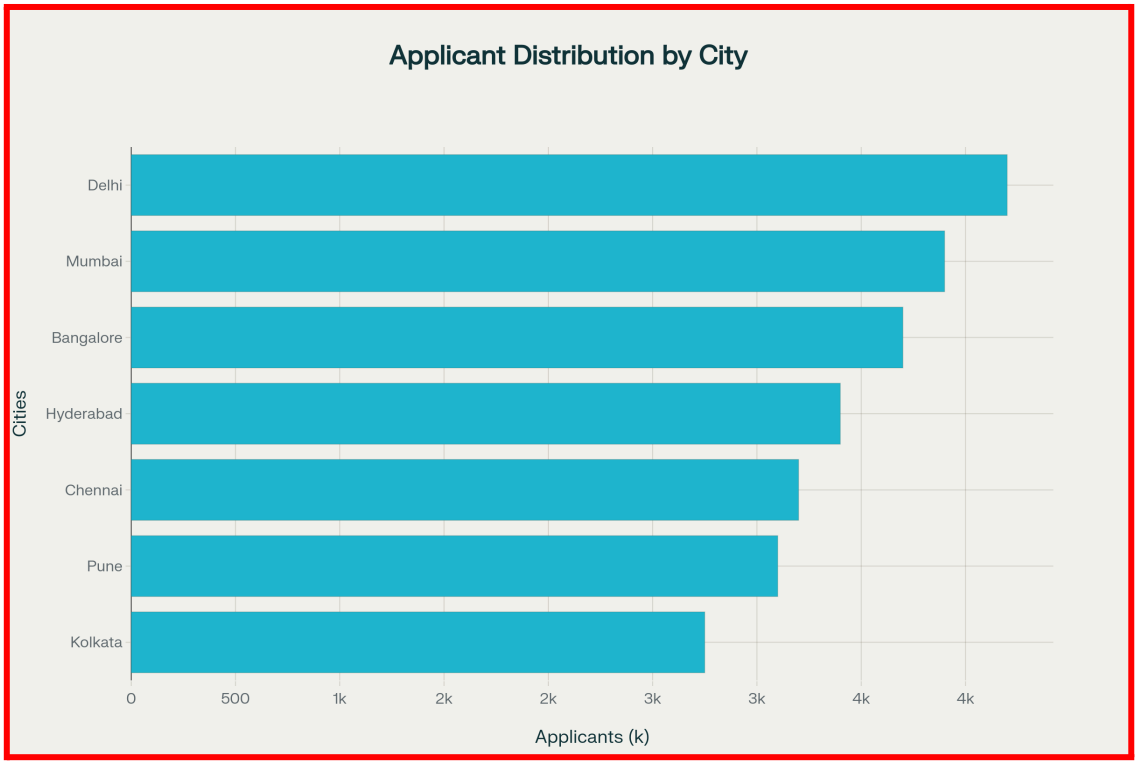
- **Bar chart of campaign status distribution in CampaignData:**



- **Scatter plot of phone number length vs outreach attempts:**



Horizontal bar chart illustrating the number of applicants from the top 7 cities in the ApplicantData, allowing you to quickly identify which urban locations contribute the most to your applicant pool.



9. Challenges Faced

Data Variation:

There is a noticeable skew in the dataset, with some metrics represented by only one or two individuals. These extreme or isolated cases create outliers and can distort overall summary statistics.

Inconsistent and Reduced Dataset:

Some variables have missing or incomplete content, such as null values or missing fields. This affects statistical reliability and complicates the analysis process.

Need for Data Normalization:

Preliminary data exploration confirms that many fields require normalization for effective cross-variable analysis.

10. Up Next

Data Reduction:

Focus next analysis on selected core variables using either unsupervised methods or by targeting groups of variables for dimension reduction and improved clarity.

Clustering and Segmentation:

Use clustering techniques to segment users and identify natural groupings, which helps in discovering patterns across applicant segments, outreach outcomes, or campaign categories.

Advanced Metric Evaluation:

Explore advanced statistical and machine learning approaches to uncover hidden patterns, predictors of success, or efficiency in outreach methods.

- Time-based & Cross-sectional Analysis: Perform time series and cohort-based analyses to track user behaviors, campaign effectiveness, or outreach success over time.

- Feature Selection and Model Building: After cleaning and reducing dimensionality, select key variables as features for building robust predictive models or dashboards.

11. Conclusion

Data Must Be Cleaned

Data quality can degrade over time due to various factors, such as merges from external sources, updates in data collection methods, or changes in user behaviors. Therefore, implementing routine checks, validation methods, and refining outlier treatment is crucial to assure the ongoing reliability and relevance of the dataset for future analysis.