Client Segmentation and Retention Analysis for Raftaar

A mid-term submission report for the BDM capstone Project

Submitted by

Name: Sujash Bharadwaj

Roll number: 23f1000914

IITM Online BS Degree Program,

Indian Institute of Technology, Madras, Chennai

Tamil Nadu, India, 600036

Contents

1. Executive summary

The Raftaar Business Data Management Midterm Submission involves an analysis of customer visit data collected from February 2, 2024, to January 2, 2025. The dataset, extracted from the Racefacer software, provides insights into customer behavior, visit frequency, and karting activity. This analysis aims to address customer retention challenges and market positioning strategies for Raftaar.

The dataset contains key variables such as age, visit frequency, last visit date, and participation in karting sessions. Initial observations indicate that 60% of visitors are one-time customers, highlighting an opportunity for improved retention efforts. By structuring data-driven marketing and engagement strategies, this project seeks to identify patterns in customer visits and develop targeted promotional efforts.

Descriptive statistics and segmentation techniques will be used to classify customers into distinct behavioral groups, allowing Raftaar to design personalized outreach initiatives. Additionally, insights from the data will contribute to optimizing event hosting strategies and enhancing long-term customer engagement.

The findings from this phase will serve as a foundation for more refined analyses in the final submission, focusing on predictive modeling, enhanced retention strategies, and business optimization techniques.

2. Proof of originality of Data

Access the evidence substantiating the originality and authenticity of the datasets utilized for the Business Data Management project can be found in this link:

clients

Data Source and clarification

clients.csv: The dataset was obtained directly from the Racefacer software, which is used by Raftaar for managing customer records and karting activity. The file was downloaded via the local server by accessing the system through the host user account.

Racefacer automatically logs customer data at the time of registration, recording key details such as:

Personal Information: Name, email, age, date of birth, phone number, country, city, postal code.

Karting Activity: Number of heats (sessions), total visits.

Registration & Engagement Details: Registration date, last visit date, email subscription status.

This dataset is sourced exclusively from Raftaar's internal system and has not been modified or obtained from external sources. The data reflects real customer interactions recorded over the period from February 2, 2024, to January 2, 2025.

This structured data provides a foundation for analyzing customer retention trends, segmentation, and engagement levels, enabling a deeper understanding of visit frequency and customer behavior.

Confirmation Documentation and Other Evidence

Link to Gdrive

MID-TERM SUBMISSION

3. Metadata and Descriptive statistics

Metadata:

Data Composition

The dataset consists of [Total Columns] attributes capturing customer details, karting participation, and engagement metrics. Key columns include:

Name, Age, Date of Birth – Basic customer identification.

Phone, Email – Contact information, with some missing values in phone numbers and emails.

Visits, Heats – Number of total visits and karting sessions completed.

Registered At, Last Visit At – Date of customer registration and last recorded visit.

Emails Accepted – Indicator of whether customers have opted in for marketing emails.

The presence of null values in Phone and Email impacts direct engagement strategies but does not affect overall visit pattern analysis.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Name	Email	Age	Date of Birth	Phone	Country	City	Postal Code	Heats	Visits	Registered at	Last Visit	Accept Emails
2	Soham Mahajan	sohammahajan007@gmail.com	25	16 Dec 1999	+919834332617	India	Pune				1/2/2025		Yes
3	Ria Malpani	rianavya@gmail.con	19	05 Nov 2005	+919309716239	India	Pune				1/2/2025		No
4	Ish Nagarkar	nikish.nagarkar@gmail.com	25	17 Oct 1999	+917744978832	India	Pune				1/2/2025		No
5	Amit Pampatwar	amit.pampatwar@cloud.com	38	16 Feb 1986	+919860453447	India	Pune				1/2/2025		No
6	Mayank Das	mayank.ngp@gmail.com	29	01 May 1995	+919767657890	India	Pune				1/1/2025		No

Data Integrity & Anomalies

Missing Values: Notably, the Phone and Email columns have significant missing entries, which may affect direct communication efforts but do not hinder visit trend analysis.

Date Formatting: All date-related fields have been standardized for consistency, ensuring accurate time-based analysis.

Visit Frequency Outliers: The Heats and Visits columns show high variability, suggesting a mix of one-time and highly engaged repeat customers. Outlier detection may be necessary to normalize analyses.

Email Opt-In Behavior: The Emails Accepted column contains a clear binary indicator, enabling segmentation for future outreach.

Descriptive Statistics

Numerical Analysis

	A	B	C	D	E	F	G	H	I	J	K
1		Count	Mean	Standard Deviation	Minimum	25th Percentile	Median	75th Percentile	Max	Unique Values	Missing Values
2	Age	20602	24.79516552	6.993890135	6	21	24	28	125	72	1
3	Heats	17294	1.748236383	4.491025131	1	1	1	2	238	63	3309
4	Visits	17081	1.38768222	1.990886266	1	1	1	1	79	42	3522
5											

1. Age

- The average customer age is 24.8 years, with the majority of customers falling between the 21st and 28th percentiles.
- The data shows 71 unique age values, suggesting a wide range of customers, primarily young adults.
- The minimal number of missing values makes this a reliable variable for segmentation.

2. Heats (Karting Sessions)

- Customers have an average of 1.75 karting sessions, but the standard deviation of 4.49 indicates substantial variability in engagement.
- While most customers complete only 1–2 sessions, outliers exist with over 200 sessions, pointing to a small group of highly active visitors.

- Around 3,309 entries are missing, which could represent customers who registered but didn't participate in a session.

3. Visits

- The average number of visits is 1.38, with a median and 75th percentile of 1, confirming that most customers are one-time visitors.
- Some customers have made up to 79 visits, indicating strong loyalty among a small segment.
- 3,522 values are missing, possibly due to incomplete logs or walk-in entries.

Insights:

These statistics highlight two clear customer segments: a large base of casual, one-time users, and a smaller group of repeat or loyal visitors. This suggests a need for personalized retention strategies aimed at converting first-time users into returning customers through targeted promotions or loyalty programs.

Categorical Analysis

The dataset includes several categorical variables such as Name, Email, Date of Birth, Phone Number, Country, City, Postal Code, and Marketing Opt-in Status. These fields provide meaningful insights into customer demographics, data completeness, and the potential for engagement and personalized communication strategies.

Name

The Name column revealed multiple repeated entries, indicating the presence of loyal or returning customers. A frequency analysis showed that the most common name appeared 14 times. Although this column is not ideal for statistical modeling, it offers a glimpse into customer recurrence patterns and can support basic loyalty tracking.

Email

Approximately 69 entries were missing email addresses, and over 1,300 entries contained placeholders such as hyphens ("-"), limiting the feasibility of email-based outreach campaigns. Most email entries were unique, which suggests that each record generally represents a distinct customer. Strengthening data validation during registration would ensure more reliable email data for future engagement initiatives.

Date of Birth

The most frequently occurring date of birth was **01 January 2000**, a strong indicator that some users may have entered default or placeholder values. Nevertheless, the column exhibited high data completeness with minimal missing entries. With appropriate cleaning, this field could be used for age-based segmentation and customer engagement strategies such as birthday rewards or targeted promotions.

Phone Number

The Phone Number column exhibited a high number of placeholder entries, including over **1,300** instances of hyphens and other generic inputs such as "0000000000". Although most numbers were unique, these invalid entries limit the potential for phone-based marketing or direct communication. Implementing validation checks at the point of data entry would significantly enhance the quality and utility of this field.

Country

The Country column was overwhelmingly dominated by entries from **India**, which accounted for over **99 percent** of all customer records. A few isolated entries came from countries such as the United States and Germany. Only one record had a missing value in this field. The data confirms Raftaar’s primarily local customer base while highlighting occasional international reach.

City

The City column showed that the vast majority of customers were from **Pune**, where Raftaar is located. Other cities such as Mumbai and Hyderabad appeared far less frequently. This confirms the business's strong local presence and indicates opportunities for hyperlocal promotional campaigns or partnerships.

Postal Code

The Postal Code column was largely complete, although it included some duplicate and placeholder values, such as hyphens. While not directly used in current segmentation, this field can support further geographic clustering when used alongside city and country data.

Marketing Opt-in Status

The Marketing Opt-in column indicates whether customers have agreed to receive promotional emails and communications. This binary field provides an immediate opportunity for segmentation. A substantial number of users opted in, making this variable highly actionable for developing targeted marketing campaigns, personalized offers, and re-engagement strategies.

4. Detailed Explanation of Analysis Process and Method

This project focuses on addressing Raftaar's primary challenges: a high proportion of one-time visitors and a lack of structured customer segmentation. To tackle these, a structured, data-driven analytical approach was adopted using behavioral variables such as visit frequency, karting session engagement, recency, and marketing opt-in status.

1. Descriptive Analysis and Data Cleaning

The analysis began with cleaning and validating the dataset. Columns were standardized, and categorical fields such as phone numbers, email addresses, and city names were reviewed for missing or placeholder entries (e.g., "-", "0000000000"). Numerical columns like Age, Visits, and Heats were evaluated using mean, median, and percentile calculations to understand customer demographics and usage patterns.

2. Behavioral Segmentation

Customers were segmented based on their number of visits and karting activity (heats). Specifically:

- Repeat Visitors were defined as customers with more than one recorded visit.
- High Engagement customers were those who completed more than three karting heats.

This allowed the identification of Raftaar's most loyal and active customers, which is essential for designing loyalty programs and personalized marketing campaigns.

3. Recency-Based Segmentation

The field Last Visit At was used to calculate Recency, defined as the number of days since the customer's most recent visit, with a reference date of January 2, 2025. Customers were then flagged as:

- Dormant: if they had not visited in over 30 days
- Recent: if they visited within the last 30 days

This segmentation is valuable for re-engagement strategies, allowing Raftaar to target dormant users with special offers or events.

4. Marketing Opt-in Segmentation

The Accept Emails column was used to create a binary segmentation of customers who have agreed to receive promotional communication. This group forms the immediate base for targeted campaigns such as newsletters, birthday coupons, and time-limited discounts.

5. Tools Used

- Google Sheets: Used for formula-driven segmentation (e.g., Recency, Repeat Visitor flags, Dormancy)
- Pivot Tables: Generated frequency summaries and insights for categorical columns (e.g., most common cities, missing email count)

- Python: Used to validate the analysis, generate segmentation summaries, and assist in preparing charts and summary data for the report

These tools provided a balance of accessibility, accuracy, and scalability, allowing deep insights without the need for overly complex modeling at this stage.

6. Justification for Approach

Given the business goal is to improve retention and engagement—not prediction—this approach is suitable. Rather than applying machine learning or statistical modeling prematurely, this method focuses on:

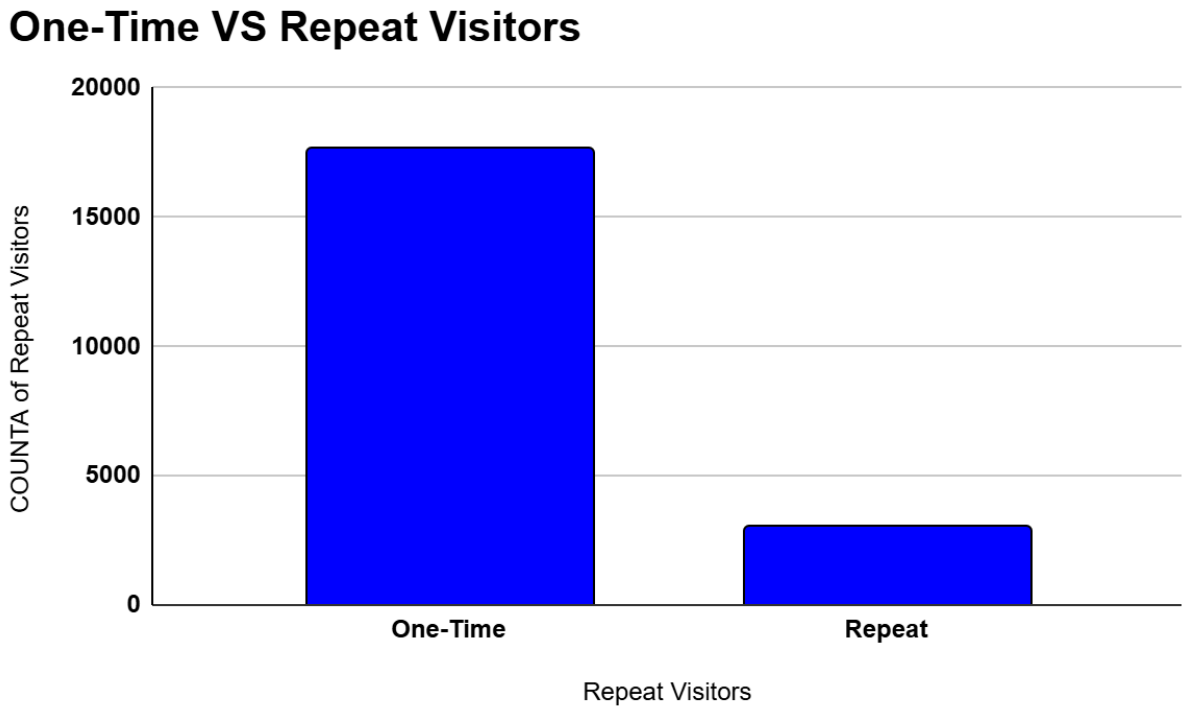
- Understanding who the customers are
- Identifying actionable segments
- Laying a solid foundation for more advanced analysis in the final submission

For the next phase, deeper techniques such as customer clustering, RFM analysis, or predictive modeling (e.g., churn likelihood) may be applied once this foundational layer is validated and business feedback is incorporated.

5. Results and Findings

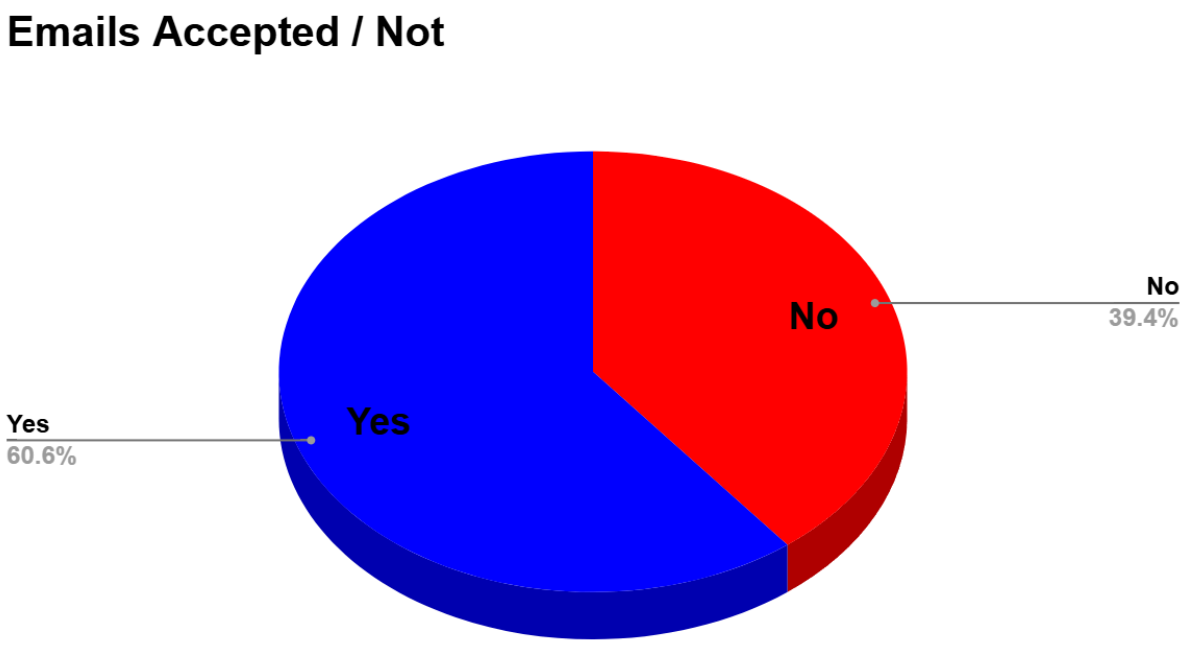
The analysis revealed several actionable insights based on customer behavior and demographic attributes. The charts below illustrate key findings derived from segmentation and frequency analysis.

1. One-Time vs Repeat Visitors



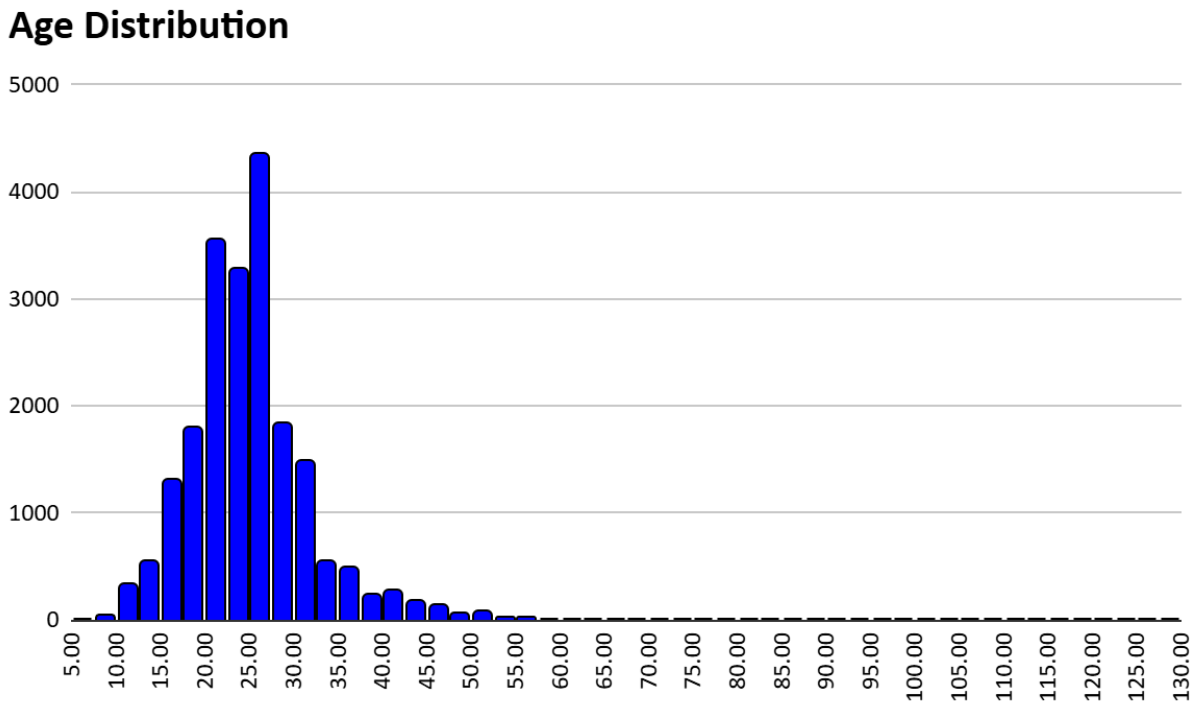
Out of 20,603 total customers, approximately 17,612 were one-time visitors, while only 2,991 returned for a second visit or more. This chart visually confirms Raftaar's major retention challenge and emphasizes the importance of developing a structured follow-up and loyalty strategy to convert first-timers into repeat customers.

2. Email Opt-In Breakdown



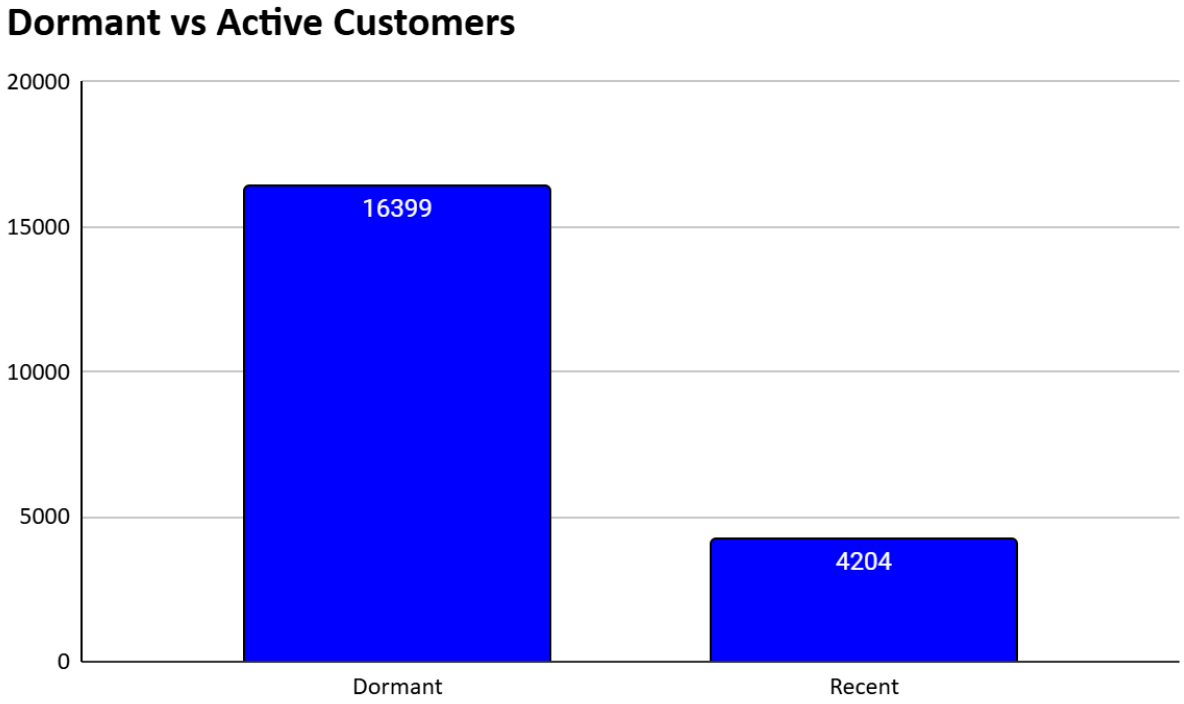
Around 60.6% of customers have opted in to receive promotional emails. This is a strong base (12,490 users) for email marketing campaigns. However, the remaining 39.4% limits the reach of digital communication, suggesting the need to improve opt-in conversion during onboarding.

3. Age Distribution



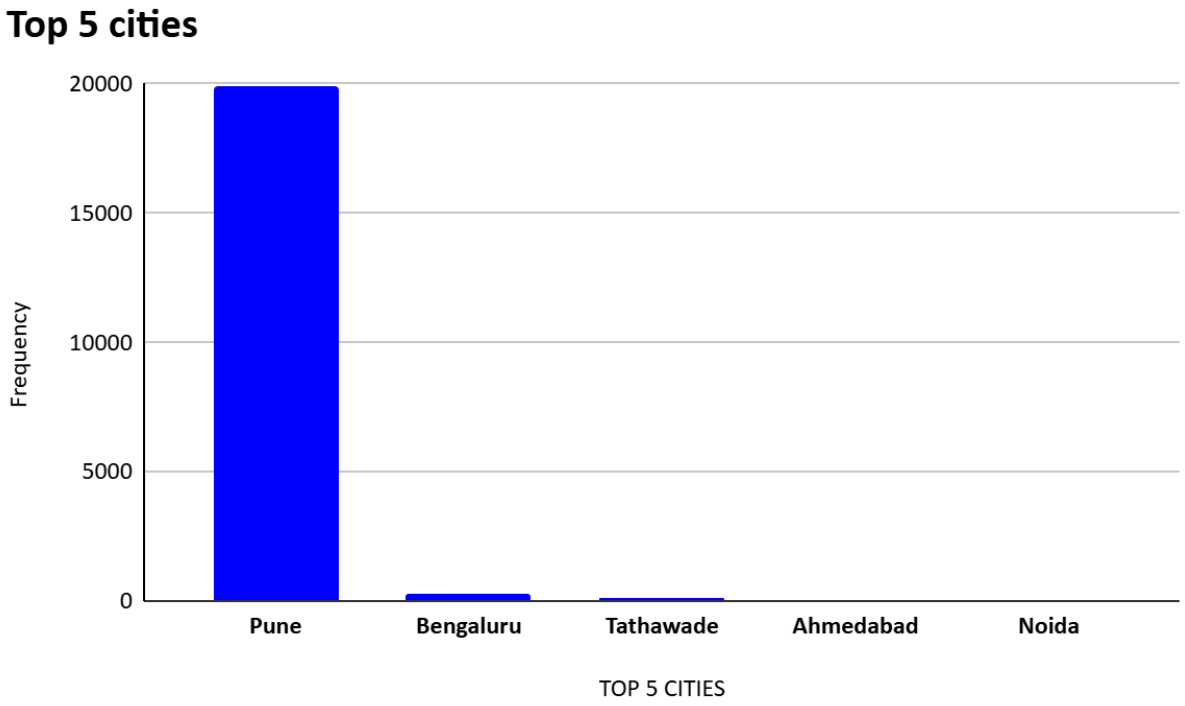
The age histogram shows that the majority of Raftaar's users are between 18 and 30 years old, with a peak around 25. This confirms a young adult demographic, aligning with Raftaar's brand as an energetic, experience-driven entertainment business. However, the presence of a few age outliers (e.g., above 70) suggests data entry errors or shared accounts.

4. Dormant vs Active Customers



A large segment (over 16,000) of customers have not returned in more than 30 days, classifying them as dormant. In contrast, 4,204 customers are still active or recently engaged. This highlights a clear need for reactivation campaigns — including discounts, events, or gamified experiences — to bring dormant users back into the fold.

5. Top 5 Cities



Pune dominates the customer base, followed by minor representation from cities like Bengaluru, Tathawade, and Noida. This supports hyperlocal marketing, targeted offers, and possible offline engagement campaigns in Pune. Geographic insights like this also validate where Raftaar's strongest word-of-mouth and digital traction lie.

Summary of Business Implications

- A large pool of dormant and one-time users provides immediate opportunity for re-engagement and retention strategies.
- Email opt-in users offer a reachable audience for targeted communication.
- Younger customers dominate the demographic, allowing Raftaar to design youth-focused experiences.
- Local customer dominance reinforces Pune as the priority geography for short-term growth and pilot campaigns.