FINAL PROJECT REPORT

WEVESTR GROWTH & BUSINESS DEVELOPMENT: DATA-DRIVEN INSIGHTS

(04/18/2025)

1. DATA COMPILATION AND EXPLORATION

➢ Data Compilation

The initial phase of this project involved data compilation, which was executed in a structured manner to ensure seamless data integration for further analysis.

- Data Acquisition: The necessary sheets from the WeVestr dataset were sourced from Google Sheets and downloaded in CSV format.
- Data Integration: A Python script was developed and executed in PyCharm to automate the compilation of all CSV files into a single Excel workbook, with each dataset organized into separate sheets for structured accessibility.
- Data Storage & Management: The compiled Excel file was successfully generated and stored within the PyCharm project directory. It was then transferred to the designated file location, ensuring proper organization and accessibility for subsequent data exploration and analysis.

This process established a well-structured, centralized data set, allowing for efficient exploration, cleaning, and visualization in the next phase

➢ Data Exploration

Following the successful compilation of the dataset, the next phase focused on data exploration, which involved examining the dataset’s structure, content, and data types to ensure its readiness for analysis.

- Loading the Compiled Dataset: The consolidated Excel file was imported into PyCharm within a dedicated script, *explore_data.py*. The pandas library was utilized to load the dataset and retrieve all available sheets within the file, ensuring structured access to the data.
- Inspecting Data Structure & Schema: A script was executed to retrieve and display the first few rows of each sheet, allowing for an initial assessment of the dataset’s structure and content. The column names, data types, and overall structure were examined to understand the dataset’s composition and potential inconsistencies.
- Data Type Validation & Consistency Check: The *dtypes* attribute of pandas was used to verify the data types of each column across all sheets. This helped identify and flag incorrect data types, missing values, and unexpected formats, ensuring data integrity before proceeding with further analysis.
- Missing Value Analysis: The dataset was analyzed for null or missing values in each sheet using *pandas' isnull().sum()* function, helping assess data completeness and identify fields requiring imputation or cleansing.
- Unique Value Distribution: Unique values were extracted for each column in every sheet, providing insights into category distributions, duplicate entries, and potential

inconsistencies. This step facilitated the detection of outliers and irregularities in categorical and numerical fields, ensuring a more refined dataset for analysis.

This foundational exploration provided critical insights into the dataset's quality and structure, forming the basis for data cleaning, transformation, and visualization in subsequent phases.

2. DATA CLEANING & PREPARATION

The second phase of the project focused on ensuring the dataset's integrity, consistency, and readiness for analysis. This involved handling missing values, removing duplicates, and compiling the cleaned dataset for further processing.

➢ **Handling Missing Values**

To ensure completeness and reliability, the dataset was analyzed for missing values and treated accordingly:

1. Missing Value Analysis:

- A Python script (missing_values_analysis.py) was executed to identify and quantify missing values in each column across all sheets.
- The percentage of missing values for each column was calculated to determine the extent of data gaps.

2. Imputation & Cleaning:

- This process was executed using the script handle_missing_values.py.
- A threshold was set to drop columns with more than 50% missing values.
- Missing values in numerical columns were filled using the mean of the respective column.
- Categorical columns were filled using the most frequent value (mode).
- Forward-fill and backward-fill techniques were applied where necessary to maintain consistency in time-series data.

➢ **Removing Duplicates**

To eliminate redundant data and maintain integrity:

1. Duplicate Identification:

- The dataset was checked for duplicate entries using the duplicated () function.

2. Duplicate Removal:

- Identified duplicate records were removed while ensuring that original, unique records were retained.
- This step was automated through the script remove_duplicates.py

Compiling Cleaned Data

After performing the necessary cleaning steps, the final dataset was prepared for analysis:

1. Integration of Cleaned CSV Files:

- All cleaned data files were stored in a designated data/ directory.
- The script compile_cleaned_data.py was used to merge these cleaned CSV files into a single structured dataset.

2. Final Storage & Export:
   - The consolidated dataset was saved in a new CSV file using save_cleaned_data_csv.py, ensuring it was formatted correctly for further analysis.

Outcome of Data Cleaning & Preparation

Improved Data Quality: Inconsistencies, missing values, and duplicates were effectively addressed, ensuring a cleaner dataset.

Structured Dataset: The final dataset was structured and formatted for easier integration into the upcoming analysis and visualization phase.

Efficient Processing: Automation scripts were implemented to streamline the cleaning process, making it reproducible for future updates.

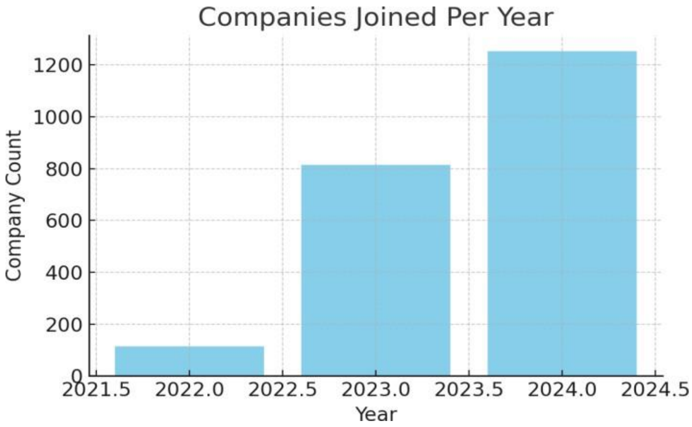
This step successfully refined the dataset, ensuring that subsequent analytics and visualization tasks are built on a robust, reliable foundation. The next phase will involve Data Analysis & Visualization, where key business insights will be derived from the cleaned dataset.

3. WEVESTR GROWTH, ONBOARDING & FEATURE ADOPTION ANALYSIS

This report presents a detailed analysis of WeVestr’s platform usage trends across company growth, onboarding sources, user activity, and feature adoption. It highlights behavioral patterns, platform engagement, and provides data-backed recommendations for product optimization and growth strategy.

1. Company Growth Overview

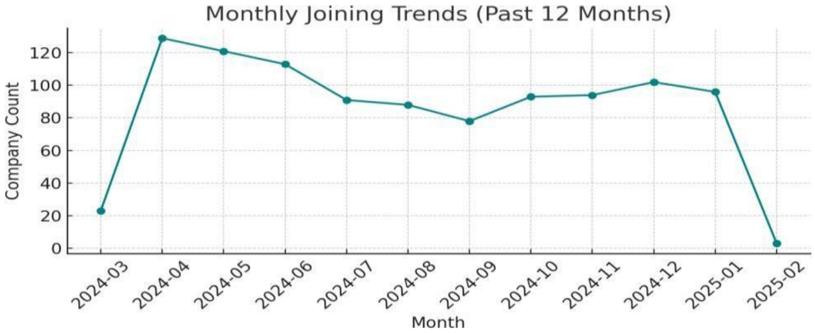
Analyzed company onboarding patterns across recent years to identify WeVestr's platform growth trajectory.



Insight: WeVestr saw consistent and strong growth from 2022 to 2024. The number of companies on board in 2024 exceeded 1200, indicating increasing market traction.

Recommendation: Capitalize on this growth trend by maintaining onboarding momentum and ensuring infrastructure scalability.

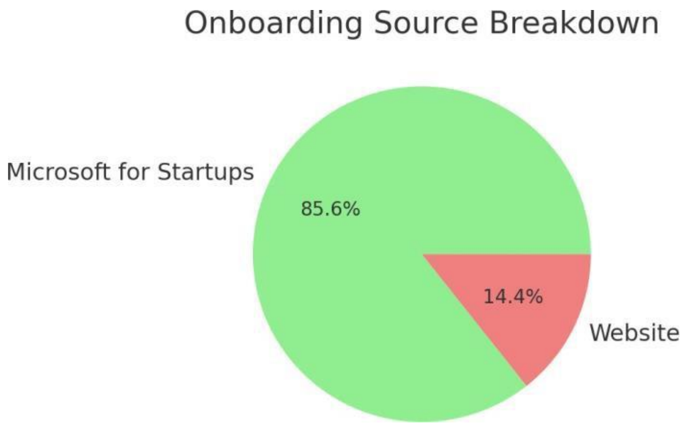
2. Monthly Joining Trends (Past 12 Months)



Insight: There was a surge in signups in April 2024 followed by a slight decline. However, steady onboarding activity continued throughout the year.

Recommendation: Investigate April campaigns to identify drivers of success. Replicate or scale these campaigns seasonally.

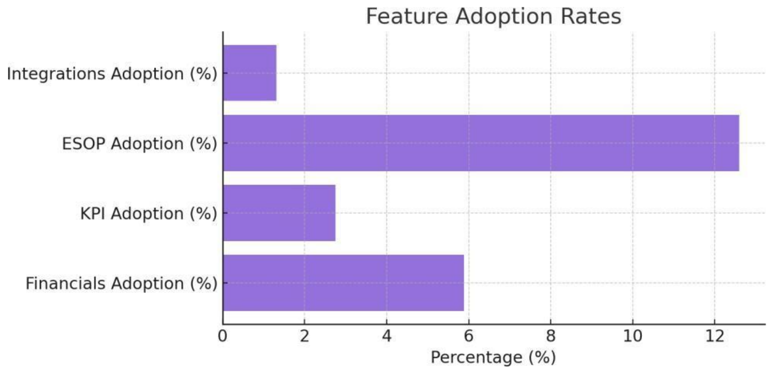
3. Onboarding Source Breakdown



Insight: The vast majority of users joined via the Microsoft for Startups program (85.6%), with only 14.4% joining via the website.

Recommendation: Improve website acquisition strategies to reduce dependency on external programs and increase direct onboarding.

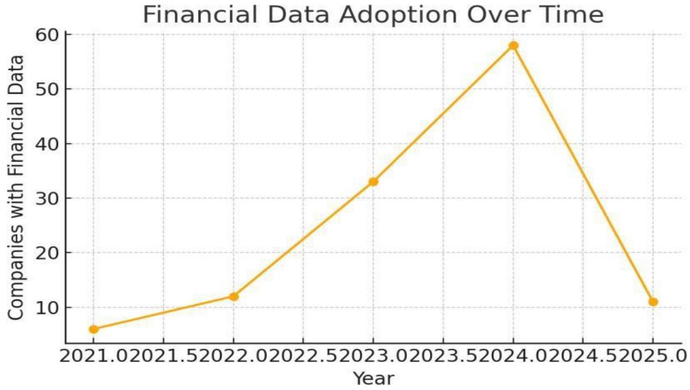
4. Feature Adoption Analysis



Insight: ESOP (12.6%) and Financials (5.9%) had higher adoption compared to KPIs (2.7%) and Integrations (1.3%). There is significant opportunity to improve feature awareness.

Recommendation: Launch targeted educational campaigns or in-app guidance to encourage use of underutilized features.

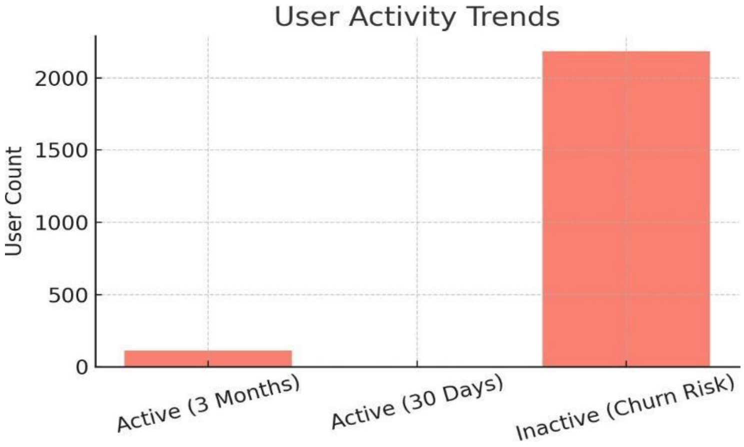
5. Financial Data Adoption Over Time



Insight: Financial data adoption has been steadily increasing since 2021, peaking in 2024 before a slight drop in 2025.

Recommendation: Identify barriers in 2025 to maintain growth and offer improved onboarding support around financials.

6. User Activity Trend



Insight: The majority of companies (95%) show no user activity in the last 3 months, indicating high churn risk. Only 4.8% were active in the past 3 months, and 0% in the last 30 days.

Recommendation: Implement retention strategies including engagement emails, in-app notifications, and support outreach for reactivation.

Conclusion

WeVestr demonstrates strong onboarding growth and promising engagement through certain features. However, there is a notable gap in user retention and a significant dependency on third-party onboarding programs. Addressing these insights through focused strategies can unlock higher platform adoption, better feature utilization, and long-term engagement.

4. ENGAGEMENT, ACTIVITY & RETENTION ANALYSIS

1. Introduction

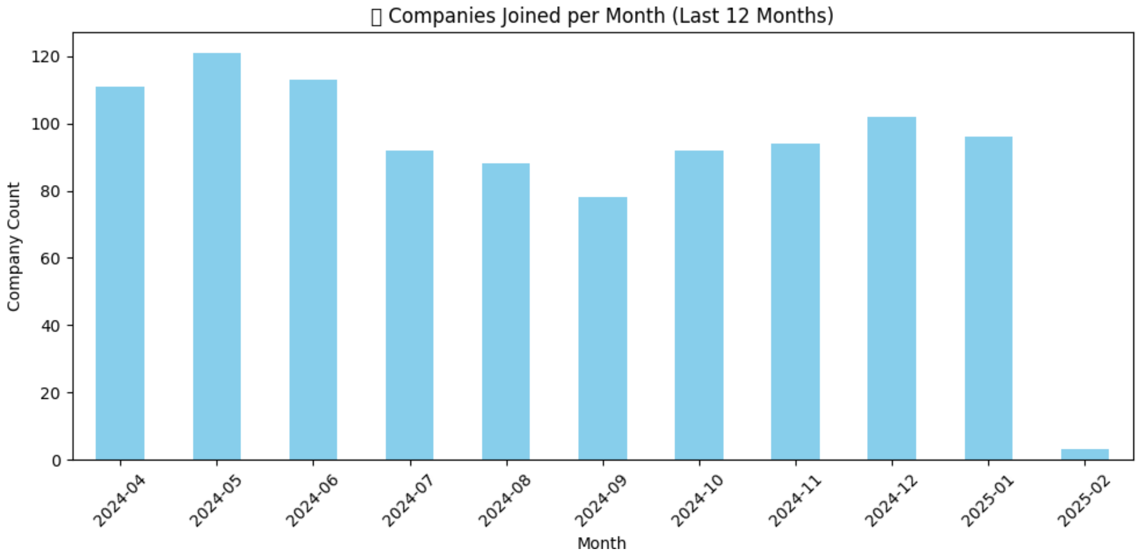
This Engagement, Activity & Retention Analysis is a crucial part of the overall WEVESTR GROWTH & BUSINESS DEVELOPMENT: DATA-DRIVEN INSIGHTS project. The goal is to analyze platform usage, engagement levels, and retention patterns of onboarded companies. By assessing metrics such as document and stakeholder engagement, monthly onboarding trends, financial updates, industry-specific activity, integration adoption, and login frequency, we can better understand user behavior, identify friction points, and drive strategic improvements for long-term growth and retention.

2. Key Questions Answered

- What is the average number of documents for companies which added at least one?
- What is the average number of stakeholders for companies which added at least two?
- What is the trend of companies joining per month in the past year?
- How many companies are actively updating their data (e.g., transactions, financial updates)?
- Which industries have the highest number of onboarded companies?
- What is the average number of documents added per industry?
- How many companies have integrated external tools (based on active integrations)?
- What is the retention rate based on login activity?

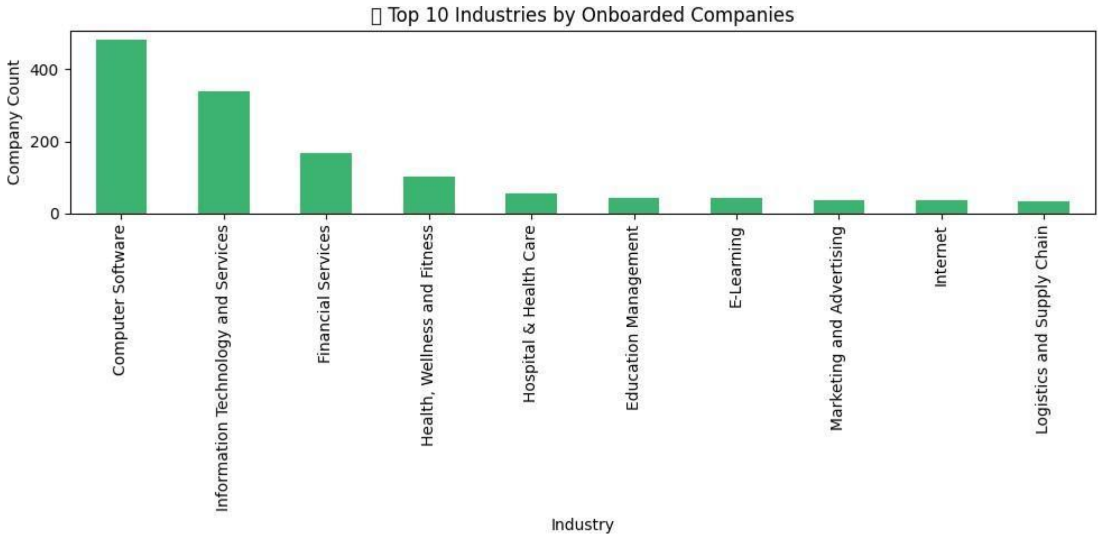
3. Key Findings & Visualizations

- The average number of documents for companies that added at least one is: 5.85
- The average number of stakeholders for companies that added at least two is: 5.30
- The chart below shows the trend of companies joining each month over the last year:

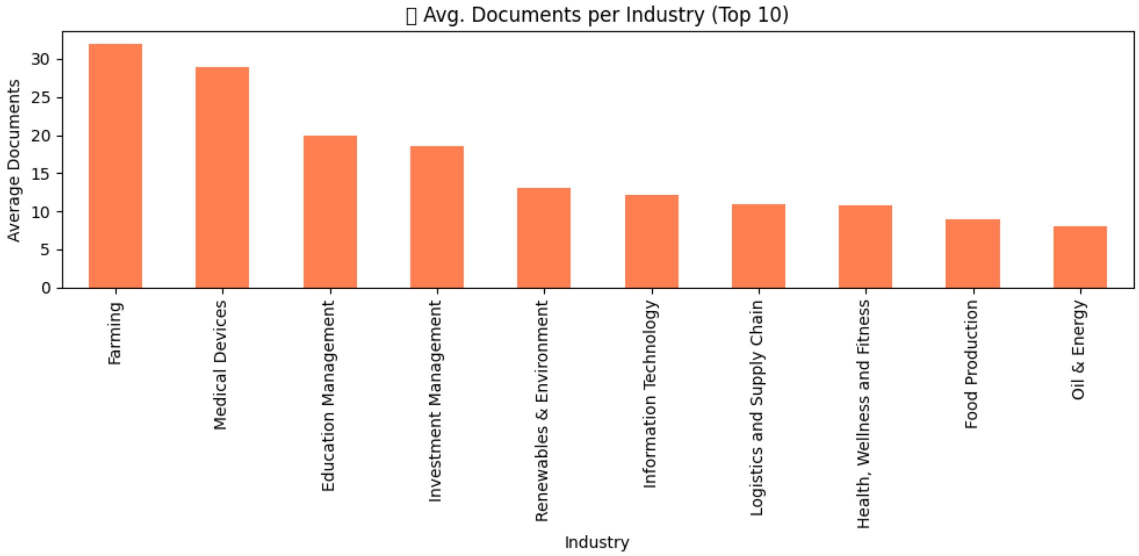


A total of 2,276 companies are actively updating their data through transactions and financial information.

The following chart highlights the top 10 industries with the most onboarded companies:



The following chart shows the top 10 industries with the highest average number of documents per company:



- A total of 30 companies have integrated external tools into the platform.
- Based on login activity (past 30 days or 3 months), the retention rate is 100%.

4. Recommendations

- Investigate industries with lower document engagement to understand platform friction or data entry limitations.
- Develop onboarding or re-engagement flows for companies with low stakeholder participation.
- Explore the use of integration partnerships to boost external tool adoption beyond the current 30 companies.
- Maintain and reinforce strategies that are contributing to the exceptional retention rate.
- Continue monitoring monthly onboarding trends to adopt strategies during declining months.
- Leverage high-performing industries such as Computer Software and Information Technology to gather product feedback and drive innovation.

Conclusion

This Engagement, Activity & Retention Analysis has highlighted key usage patterns, areas of strength, and improvement opportunities within the WeVestr ecosystem. The insights gained here support a data-informed strategy to optimize user engagement, enhance product adoption, and ensure scalable growth.

STRATEGIC BUSINESS RECOMMENDATIONS

1. Appointing a Dedicated Data Steward / Manager

Across the dataset, inconsistencies in format, missing values, and outdated entries were frequent. This significantly impacted analysis accuracy and delayed insight generation.

Recommendation:

Appoint a Data Steward responsible for managing database hygiene. This role would:

- Enforce structured data entry standards
- Schedule regular data updates from internal teams
- Conduct periodic audits for quality assurance

Outcome: Improved data reliability and faster decision-making.

2. Launch Targeted Re-engagement Campaigns

95% of users had no login activity in the last 3 months, indicating high churn risk.

Recommendation:

Deploy a multi-channel retention strategy, including:

- Triggered email reminders for inactive users
- In-app nudges encouraging reactivation
- Value-based content demonstrating new feature benefits

Outcome: Improved retention and renewed user engagement.

3. Diversify Onboarding Channels

85.6% of companies joined via Microsoft for Startups, only 14.4% via the website.

Recommendation:

Reduce reliance on third-party onboarding by:

- Enhancing SEO and landing page experiences
- Launching referral or partner programs
- Investing in organic growth via content marketing

Outcome: Sustainable pipeline of self-registered users.

4. Strengthen Feature Onboarding Experience

Adoption of advanced features (e.g., KPI, Financials, Integrations) remains low.

Recommendation:

Introduce interactive onboarding workflows such as:

- Product tours with feature walkthroughs
- Tooltips contextual to user actions
- Help center upgrades with quick-start templates

Outcome: Higher feature usage and reduced time-to-value.

5. Leverage High-Performing Industries

Industries like Computer Software and Information Technology show strong engagement and document interaction.

Recommendation:

Prioritize these segments for:

- Beta testing new features
- Testimonials and case studies
- Product feedback sessions

Outcome: Industry-focused innovation and advocacy loops.

6. Automate Trend Monitoring

Company signups peak in specific months (e.g., April 2024).

Recommendation:

Automate monthly trend detection to:

- Identify repeatable campaign patterns
- Align product updates or offers with demand spikes

Outcome: Proactive planning for seasonal growth.

7. Boost Data Integration Capabilities

Only 30 companies have connected external tools.

Recommendation:

Expand integrations with popular tools (e.g., QuickBooks, HubSpot, Google Sheets) and promote:

- Plug-and-play integration libraries
- API access documentation
- Partnership outreach with SaaS tools

Outcome: Increase stickiness and cross-functional adoption.

8. Institute Industry-Specific Playbooks

Engagement levels vary significantly by industry.

Recommendation:

Create industry-specific onboarding and successful playbooks to:

- Guide users through relevant setup steps
- Showcase tailored use cases

Outcome: Personalized experiences that drive activation.

9. Implement Data Governance Framework

Data gaps and duplication signal the need for structured governance.

Recommendation:

Adopt a lightweight data governance model:

- Define naming conventions, required fields, and update cycles
- Set up validation scripts to catch inconsistencies

Outcome: Minimized downstream data cleaning and reporting errors.

ANOMALY & DISCREPANCY ANALYSIS REPORT

OBJECTIVE

To investigate and explain the conflicting insights discovered in two stages of the project:

- Growth & Onboarding Analysis reported: “95% of companies showed no user activity in the last 3 months.”
- Engagement & Retention Analysis reported: “100% of companies had login activity in the past 3 months and 30 days.”

These contradictory results prompted a detailed anomaly investigation.

ROOT CAUSES OF ANOMALY & DISCREPANCY

The root cause lies in how the data was sourced and interpreted.

1. Missing Login-Related Fields in Core Dataset

- The "Copy of WV - RAW" sheet does not contain any login-related columns like:

  "last login(date, any user)"

  "last login(date, admin user)"

  "Login times last 3 months"

  "Login times last 30 days"
- All login-related data exists only in the "Copy of WV - Active users" sheet.

Impact: Retention measured using only the active users sheet gave a false 100% result, ignoring the full population.

2. No Unique Identifier for Merging

- There was no consistent primary key (like a user ID) to link records across sheets.
- Even Company and Email fields differ in casing (upper and lower case), format, or consistency, whitespace, making merging unreliable without preprocessing.

3. Two Sheets Used Independently

- Each analysis used its own isolated dataset, without integrating data sources. This caused insights to reflect only partial realities.

4. Missing Login Data

- When the two sheets were merged and cleaned, we found: 2,254 out of 2,296 users (98%) had no login activity recorded.

5. Inconsistent Tracking Systems

- The RAW sheet seems to be more comprehensive, but contains missing or malformed fields (e.g., blank login dates).
- The "Active Users" sheet may have been exported manually or sampled from a CRM, and may not represent the full user base.

RESOLUTION STRATEGY

The retention analysis is skewed due to sampling bias — it analyzes only those who were already active. To get a realistic retention picture, the following was performed:

- We merged the two original data sheets:
  "RAW Data Sheet" (which includes company details and onboarding info).
  "Active Users Sheet" (which contains user login records).
- We merged both sheets using cleaned primary email values.
- We treated users with no login data as inactive.
- Recalculated retention rates across the full company base.

METRICS AFTER CORRECTION

- Total companies analyzed: 2,296
- Users missing login activity data: 2,254
- Users logged in during the last 3 months: 42
- Users logged in during the last 30 days: 29
- Actual Retention rate (last 3 months): 1.83%
- Actual Retention rate (last 30 days): 1.26%

KEY INSIGHT

- The 100% retention rate initially reported was a sampling illusion.
- Only 1.26%–1.83% of companies had logged in recently — confirming significant disengagement and highlighting an urgent need for better data governance.

BUSINESS RECOMMENDATIONS TO BRIDGE DATA GAPS

1. Assign a Dedicated Database Manager

- A responsible data steward should oversee data consistency, especially for critical identifiers like email addresses. This person should be responsible for cleaning, updating, and validating data regularly.

2. Unify User Tracking System

- All login data must be logged into a single unified system accessible in the core user dataset. Create a Master User Table and combine all user activity, company details, and engagement data into one standardized sheet.

3. Ensure Full Data Coverage

- All users should have login activity tracked — not just a few. Analytics should be based on the entire user base, not partial datasets.

4. Use Consistent User Identifiers

- Standardize all email addresses or assign a unique user_id for merging across data sources.

5. Reframe Future KPIs

- All engagement metrics should reflect the full population — not just active users.

6.Improve Data Collection Processes:

- Investigate why 98% of the users are missing from login records. If external systems are used (SSO, third-party logins), ensure those logs are integrated.

7. Perform Regular Audits & Automate Data Validation:

- Conduct weekly or monthly audits to verify the consistency and completeness of all data logs (especially login-related metrics). Set up scripts that routinely scan for: Missing fields, Outdated login records or Broken links between tables.