

Customer Sign-Up Behaviour & Data Quality Audit

1. Business Scenario

The manager of a fast-growing SaaS company offering tiered subscription plans asked for a data quality audit and insights into user acquisition trends, using recent customer sign-ups and support tickets datasets. This report is intended to help the Marketing and Onboarding teams optimise their campaigns and engagement workflows.

They were particularly interested in:

- Identifying where the data may be inaccurate or incomplete;
- Understanding how users are signing up and which plans they're choosing;
- Assessing marketing opt-in behaviour and demographics.

2. Datasets

The datasets that were used for this report are:

- **customer_signups.csv** – Contains information about the users who signed up to use the service. Each row represents a single customer and includes details such as customer's id, the signup date, their contact, demographic, and geographic information (such as name, email, gender, region, and age), the source through which customers found the service, the service plan they are using, and whether the customer opted into marketing communications.
- **support_tickets.csv** – Contains records of customer support interactions. Each row represents a support ticket and includes data such as the ticket's id, the customer's id, the ticket's date, the type of issue reported by the customer, and whether the issue was resolved or not.

3. Data Cleaning Summary

To ensure accurate and relevant analysis, the data must be clean. Data that hasn't been cleaned properly can result in biased or even wrong results. The following data cleaning processes were applied:

- **Data type conversion** – The date columns were initially stored as 'object'. Changing it to 'datetime' was needed to ensure flexibility when it comes to analysis operations (such as filtering, sorting, and performing extractions or calculations).
- **Duplicate rows handling** – There were two customers with identical IDs, but different data. Since this was not a duplicate, both rows were removed. Assigning two different customers the same ID can lead to inaccurate results.
- **Standardisation of text values** – Some columns had inconsistent data entries, such as different text formats, input data types, or misspellings:
 - '?' in *source* -> flagged as missing value (np.nan);
 - 'basic' in *plan_selected* -> 'Basic';
 - 'PREMIUM' in *plan_selected* -> 'Premium';
 - 'UnknownPlan' in *plan_selected* -> flagged as missing value;
 - 'PRO' in *plan_selected* -> 'Pro';
 - 'prem' in *plan_selected* -> 'Premium';
 - 'Nil' in *marketing_opt_in* -> flagged as missing value;
 - 'male' in *gender* -> 'Male';
 - 'FEMALE' in *gender* -> 'Female';
 - '123' in *gender* -> flagged as missing value (unlikely to be a gender, so it wasn't set as 'Other');
 - 'thirty' in *age* -> '30';
 - 'unknown' in *age* -> flagged as missing value;
 - '206' in *age* -> outlier, flagged as missing value (not a valid human age).



Figure 1: Unique values in each column, before (up) and after (down) standardisation

- **Missing values handling** – The customer_signups dataset had a total of 124 missing values. Entirely removing them would result in biased results, or even misrepresentation. They were handled as:
 - Dropping the rows with missing values for identifiers and columns that are crucial for analysis and can't be completed: customer_id, email, name, signup_date;
 - Imputing with mode values for missing values in categorical columns: source, region, gender, plan_selected, marketing_opt_in;
 - Fill with the median value for the age column.

4. Key Findings & Trends

The user acquisition strategy appears stable and effective, but also flat. In most weeks, the number of sign-ups was 7, and the least it dropped to was 4.

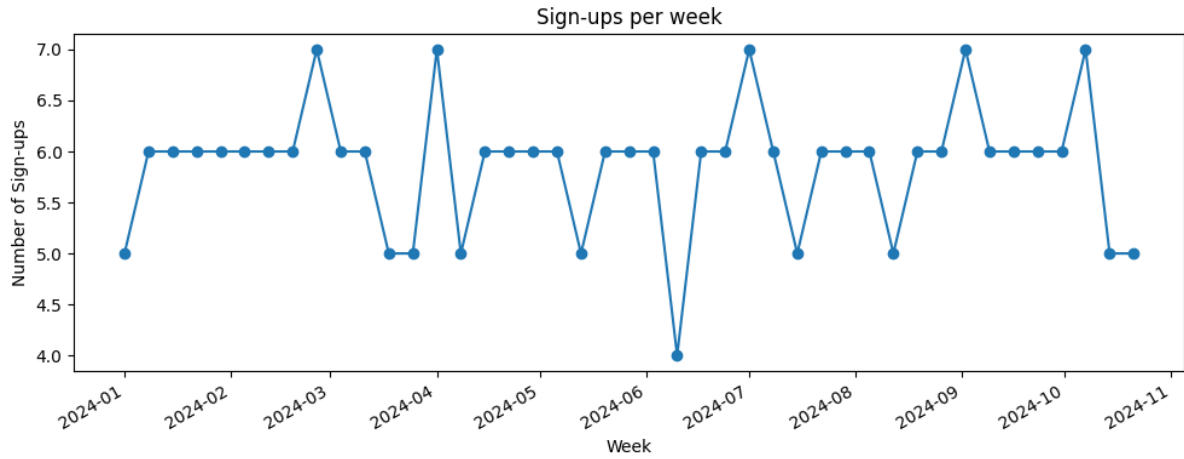


Figure 2: Sign-ups per week chart

The most sign-ups are brought by YouTube, and the most selected plan is Premium, chosen mainly by customers from the North. On the other end, the least sign-ups are brought by LinkedIn, the least selected plan is Basic, and the region with the least amount of customers is Central.

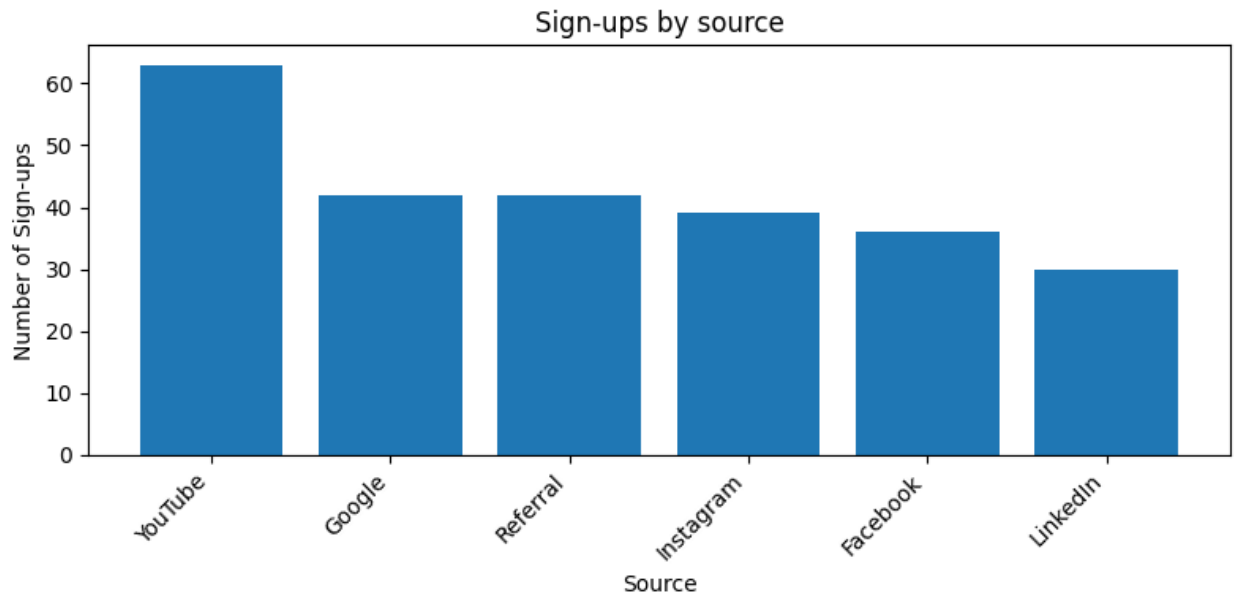


Figure 3: Sign-ups by source chart

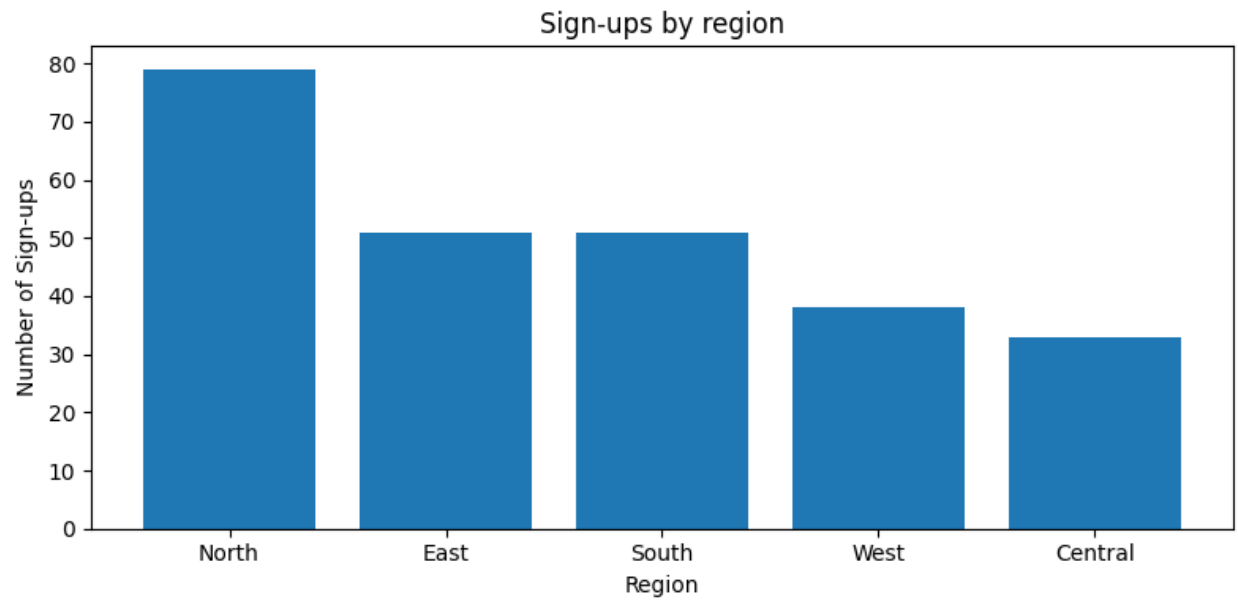


Figure 4: Sign-ups by region chart

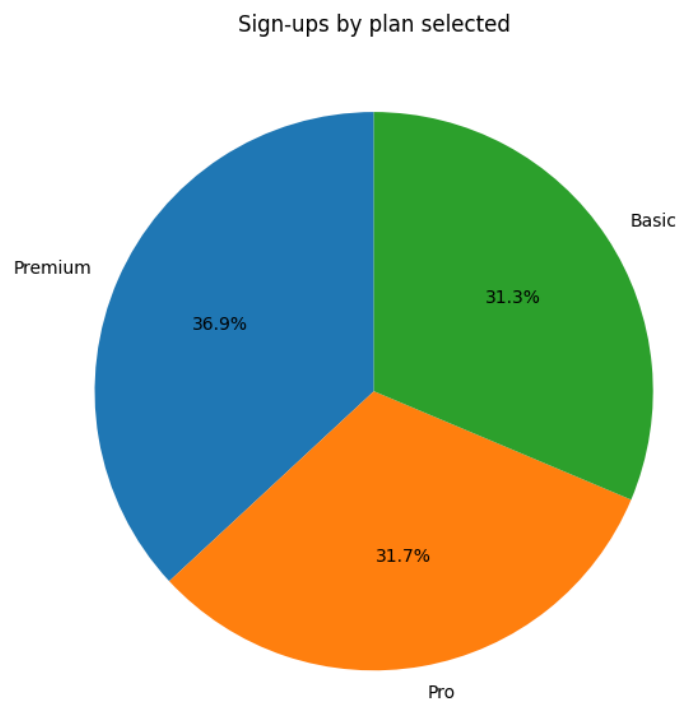


Figure 5: Sign-ups by plan selected chart

Female customers are the most likely to opt in to marketing, while the customers who are least likely to do so are non-binary.

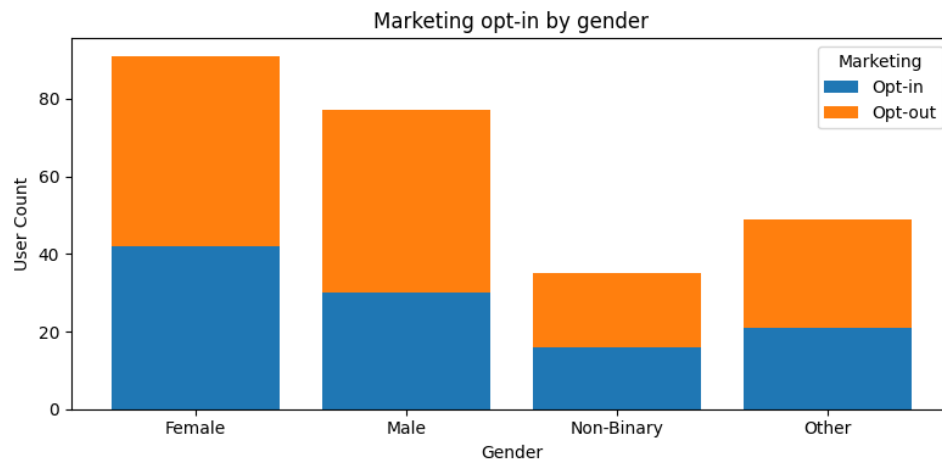


Figure 6: Marketing opt-in by gender chart

The median age of the service's customers is 34. The youngest customer is 21, while the oldest is 60.

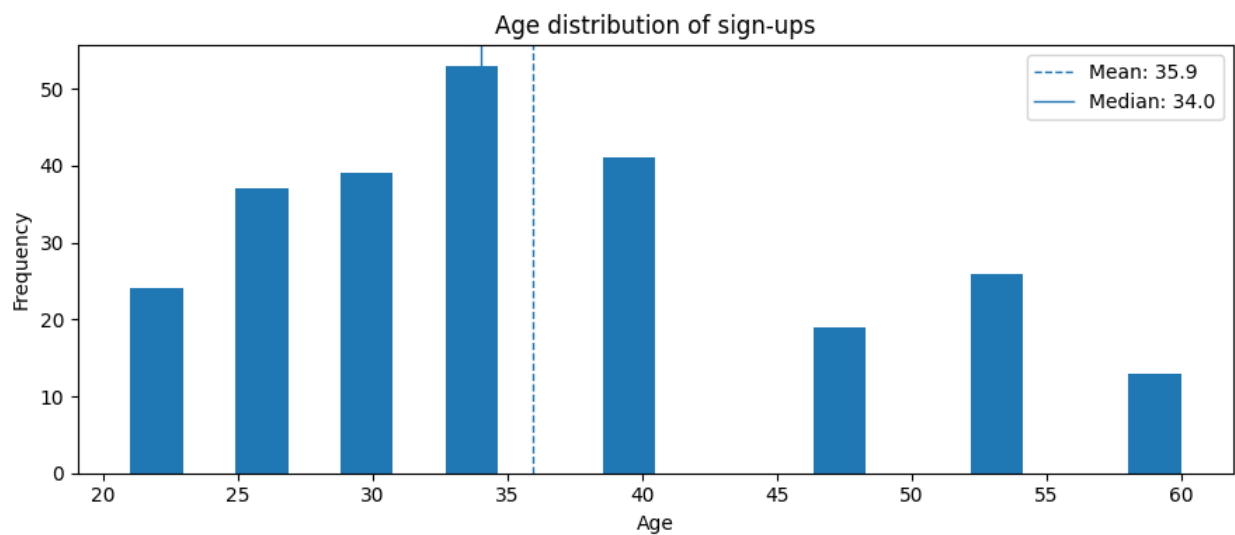


Figure 7: Customers age summary chart

5. Business Questions & Answers

Question 1: Which acquisition source brought in the most users last month?

Answer: The acquisition source that brought in the most users last month is Google, which brought in 7 users.



Figure 8: Acquisition that brought in the most users last month

Question 2: Which region shows signs of missing or incomplete data?

Answer: Central region has the least amount of data compared to the other regions.

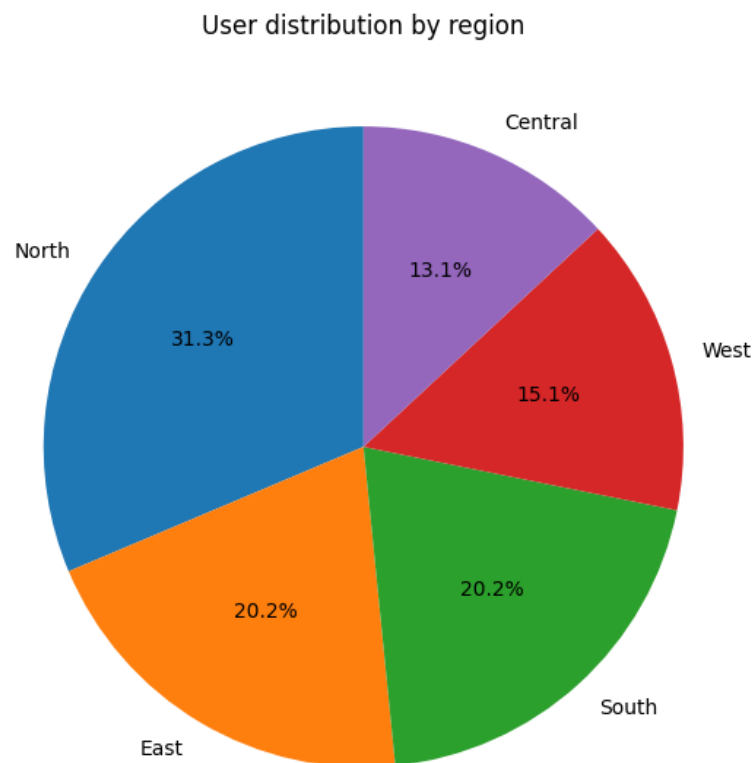
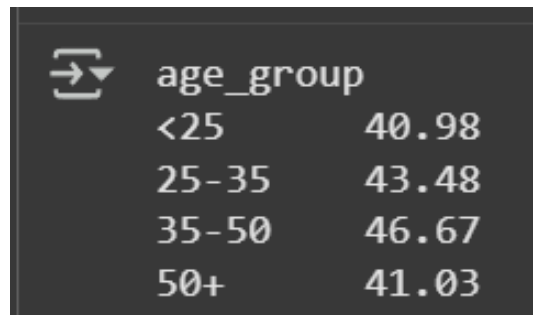


Figure 9: User distribution by region chart

Question 3: Are older users more or less likely to opt in to marketing?

Answer: No, older and younger users opt in at the same rate. The correlation index between age and marketing opt-in is ~ 0.005 , which means that there is no relevant relationship between the age and the decision to opt-in for marketing.



age_group	
<25	40.98
25-35	43.48
35-50	46.67
50+	41.03

Figure 10: Customer opt-in based on their age

Question 4: Which plan is most commonly selected, and by which age group?

Answer: The most popular plan is Premium, mainly selected by users aged 25-35.

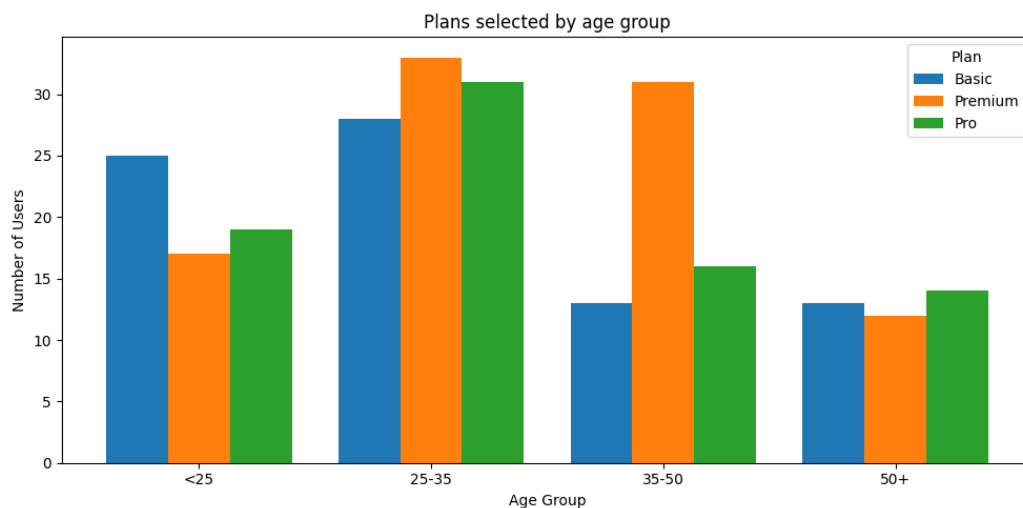


Figure 11: Plan selected by age group chart

Question 5: Which plan's users are most likely to contact support?

Answer: Pro plan's users are most likely to contact support, with a rate of 27%.

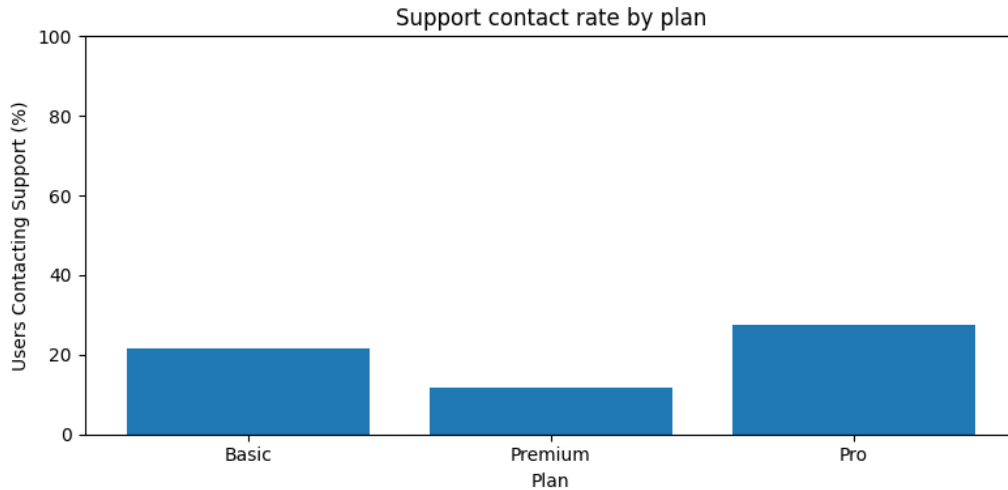


Figure 12: Support contact rate by plan selected chart

6. Recommendations

Based on the previous findings, the following suggestions would help increase the service's popularity and improve the overall customer experience:

Improve the data collection process. The data collection process is not reliable, especially in fields like region, plan selected, and gender. There were 124 missing entries. The lack of correct data could result in biased analysis due to underrepresentation or segmentation.

customer_signups dataset structure:

	DataType	Count of Missing Values	% of Missing Values
customer_id	object	2	0.666667
name	object	9	3.000000
email	object	34	11.333333
signup_date	object	2	0.666667
source	object	9	3.000000
region	object	30	10.000000
plan_selected	object	8	2.666667
marketing_opt_in	object	10	3.333333
age	object	12	4.000000
gender	object	8	2.666667

Figure 13: customer_signups dataset structure and information

A good way to fix this is to consider form and input validations, to prevent entering values such as “206” as age, “thirty” instead of “30”, and so on.

Marking fields as required so the important data is collected is also a great option.

Adding dropdown menus or selection buttons in the sign-up form instead of free-text input will prevent getting values such as “prem” instead of “Premium”.

Experiment with campaigns, referral programs, or targeting to increase weekly sign-ups beyond the 4-7 sign-ups per week plateau.

Review the onboarding experience and documentation for the Pro plan users, as most of them needed to contact for support.

Consider running campaigns in the West and Central regions to increase regional balance.

Google Colab link:

<https://colab.research.google.com/drive/1Z3g8geKlOLZII775fJkfLUT91183M34P?usp=sharing>