

Rapid Scale Customer Sign-Up Data Quality Audit & Insights

1. Introduction

This report presents a data quality audit and user acquisition analysis for Rapid Scale, a fast-growing SaaS company. The dataset reviewed contains customer sign-up records, including customer demographics, subscription plans, marketing opt-in status, and acquisition sources. The objective is to assess data accuracy, identify trends, and provide actionable recommendations to support the Monthly Business Review (MBR) and guide marketing and onboarding strategies.

2. Data Cleaning Summary

The dataset required extensive cleaning to ensure reliability. Key actions included:

- a. Changed the data types of all respective columns
- b. Removed duplicate customer IDs.
- c. Handled missing values:
 - Customer IDs: Replaced blanks with sequential IDs.
 - Names: Filled missing entries with 'Unknown' and standardized capitalization.
 - Emails: Standardized format, converted to lowercase, replaced missing/invalid with 'unknown@mailhub.org'.
 - Regions: Replaced blanks with 'North' because it was the most.
 - Source: Trimmed and replaced blanks and ?? with 'YouTube'
 - Signup dates: Standardized multiple date formats using 'regex', filled missing with most common date.
 - Ages: Converted to numeric, replaced outliers (206) and missing with median, converted to integers.
 - Marketing opt-ins: Replaced Nil and None with "No" and blanks with "Yes".
 - Gender: Standardized categories, replaced blanks with 'Male' because it was most and invalid codes, that is 123 with Female.
 - Plan selected: Corrected inconsistent labels, that is Prem to Premium, basic to Basic, PRO to Pro and assigned Basic and Pro randomly for blanks at 50%/50%

```
customer_id    name    email    signup_date \
0  CUST00000    Joshua Bryant    unknown@mailhub.org    2024-02-02
1  CUST00001    Nicole Stewart    nicole1@example.com    2024-01-02
2  CUST00002    Rachel Allen    rachel12@example.com    2024-01-03
3  CUST00003    Zachary Sanchez    zachary3@mailhub.org    2024-01-04
4  CUST00004    Unknown    matthew4@mailhub.org    2024-01-05
5  CUST00005    John Gonzales    john5@mailhub.org    2024-01-06
6  CUST00006    Crystal Mason    crystal6@mailhub.org    2024-01-07
7  CUST00007    Michael Bailey    michael7@mailhub.org    2024-01-08
8  CUST00008    Bianca Morris    bianca8@example.com    2024-01-09
9  CUST00009    Cindy Anderson    unknown@mailhub.org    2024-01-10
10 CUST00010    Kendra Gill    kendra10@mailhub.org    2024-01-11
11 CUST00011    Lisa Miller    lisa11@mailhub.org    2024-01-12
12 CUST00012    Sarah Fletcher    sarah12@example.com    2024-01-13
13 CUST00013    Michael Cole    michael13@mailhub.org    2024-01-14
14 CUST00014    Christopher Stevens    christopher14@example.com    2024-01-15
15 CUST00015    Anthony Anderson    anthony15@example.com    2024-01-16
16 CUST00016    Gary Walters    gary16@inboxmail.net    2024-01-17
17 CUST00017    Patty Paul    patty17@inboxmail.net    2024-01-18
18 CUST00018    Tara Harmon    unknown@mailhub.org    2024-01-19
19 CUST00019    Yvonne Parker    yvonne19@inboxmail.net    2024-01-20

source    region    plan_selected    marketing_opt_in    age    gender \
0  Instagram    North    Basic    No    34    Female
1  LinkedIn    West    Basic    Yes    29    Male
2  Google    North    Premium    Yes    34    Non-Binary
3  YouTube    North    Pro    No    40    Male
4  LinkedIn    West    Premium    No    25    Other
5  Facebook    South    Premium    No    34    Other
6  YouTube    North    Premium    Yes    40    Male
```

3. Key Findings & Trends

- Weekly sign-up trends indicate fluctuating acquisition activity, with identifiable peaks. This date, “2024-02-04” had the highest sign ups of 14 customers.
- YouTube emerged as the top acquisition source.
- Premium is the most selected marketing plan.
- Marketing opt-in rates are higher among 36-46 age group and opted for Yes.
- The 'North' region shows the highest sign-ups, partly due to missing-region imputation.

```
Weekly Sign-Ups:
signup_date
2024-01-07    6
2024-01-14    7
2024-01-21    7
2024-01-28    7
2024-02-04   14
2024-02-11    7
2024-02-18    7
2024-02-25    7
2024-03-03    7
2024-03-10    7
2024-03-17    7
2024-03-24    6
2024-03-31    6
2024-04-07    7
2024-04-14    7
2024-04-21    7
2024-04-28    7
2024-05-05    6
2024-05-12    7
2024-05-19    7
2024-05-26    7
2024-06-02    7
2024-06-09    6
...
max          60.000000
Name: age, dtype: float64
Null Count in Age:
0
```

The screenshot shows a Jupyter Notebook window titled 'customer_signups_analysis'. The notebook is open to a cell containing the following output:

```
... Top acquisition source last month: YouTube
```

Missing region count: 0
Unique region values:

region	count
North	94
East	61
South	59
West	45
Central	39

Name: count, dtype: int64

Marketing Opt-in by Age Group:

age_group	marketing_opt_in	No	Yes
<25		0.552632	0.447368
25-35		0.522936	0.477064
36-46		0.500000	0.500000
47-60		0.507937	0.492063

Plan by Age Group:

age_group	plan_selected	Basic	Premium	Pro
<25		30	22	24
25-35		36	37	36
36-46		11	25	14
47-60		20	21	22

4. Business Question Answers

- Data Accuracy: Issues included inconsistent date formats, blank regions, and invalid categorical values.
- User Acquisition: Most users sign up via YouTube, with Premium being the most popular plan.
- Marketing Opt-In Behavior: Higher opt-in rates occur among older age groups (36-60)

5. Recommendations

- Enforce stricter data validation at the point of entry to prevent blanks and invalid codes.
- Focus marketing efforts on strong-performing acquisition channels like YouTube.
- Develop targeted campaigns for younger age groups to improve marketing opt-in rates.

6. Data Issues or Risks

One significant data quality issue is the lack of complete regional data. Replacing blanks with 'North' may distort regional performance metrics. Future data collection processes should require mandatory region input, validated against a predefined list.