UPTRIAL DATA ANALYST INTERNSHIP

WEEK 1: Analyzing SaaS User Behavior: A Data Quality Project with Pandas

Table of Contents

1. Introduction

This report analyses customer acquisition and data integrity for Rapid Scale, a SaaS company with tiered subscriptions. The Business Intelligence team conducted a data quality audit and behavioural analysis using two datasets: 300 customer sign-ups and 123 support tickets.

The analysis aimed to:

1. Assess data quality by identifying inaccuracies and completeness issues.
2. Uncover patterns in acquisition, channel effectiveness, and plan preferences.
3. Provide actionable intelligence to optimize marketing campaigns and enhance onboarding workflows.

2. Data Cleaning Summary

The datasets underwent comprehensive cleaning to establish a reliable foundation for analysis, addressing critical data quality issues through systematic approaches.

2a) Missing values

```
MISSING VALUES - CUSTOMER SIGNUPS:
                Missing Count  Missing %
customer_id                 2       0.67
name                        9       3.00
email                      34      11.33
signup_date                 2       0.67
source                      9       3.00
region                     30      10.00
plan_selected               8       2.67
marketing_opt_in            9       3.00
age                        12       4.00
gender                      8       2.67

==================================================
MISSING VALUES - SUPPORT TICKETS:
            Missing Count  Missing %
ticket_id               0        0.0
customer_id             0        0.0
ticket_date             0        0.0
issue_type              0        0.0
resolved                0        0.0
```

Figure 1: Missing values Distribution

The analysis revealed varying levels of missing data across columns, each requiring tailored treatment based on business criticality.

- Region data gaps (10%) were categorized as "Unknown" to preserve analytical integrity

- Missing age values (4%) were replaced with the median age of 34 years to maintain dataset consistency
- Email addresses (11.4%) were retained as null values to avoid misrepresenting communication capabilities

2b) Standardization of inconsistent values

Data standardization was implemented across multiple fields to ensure consistency

- Plan Selection: Consolidated 8 variations (Basic, PREMIUM, Pro, Premium, PRO, prem) into only basic and premium.
- Gender Categories: Standardized 6 formats (Male, male, FEMALE, Female, Non-Binary, Other) to male, female, other.
- Boolean Conversion: Transformed Yes/No text to True/False for marketing_opt_in and resolved columns
- Invalid Values: Replaced UnknownPlan, 123, and ?? with NaN for data integrity
- Date Format: Standardized text dates (2nd February 2024) to DD/MM/YYYY format

2c) Outlier detection

The analysis identified and corrected an impossible age value of **206 years** for customer CUST00204 by replacing it with the median age of 34 years. This approach preserved data integrity while preventing analytical distortion.

2d) Data Type Conversions

Initial assessment revealed all columns were stored as object datatype, requiring systematic conversion of key columns to appropriate formats for accurate analysis:

- **Signup dates** were converted from **text to datetime** format, enabling time-series analysis of acquisition trends
- **Marketing opt-in** and support resolution status were converted from text (Yes/No) to **boolean** values (True/False) for logical operations
- **Age** values were standardized to **float64** format, supporting statistical analysis and demographic profiling

3. Key Findings & Trends

3a) Sign-ups per week grouped by signup_date

SIGN-UPS PER WEEK:

week	customer_id
1	6
2	7
3	7
4	7
5	8
6	7
7	7
8	7
9	7
10	7
11	7
12	6
13	6
14	7
15	7
16	7
17	7
18	6
19	7
20	7
21	7
22	7
23	6
24	6
25	7
26	7
27	7
28	7
29	6
30	7
31	7
32	6
33	7
34	7
35	7
36	7
37	7
38	7
39	7
40	7
41	7
42	6
43	6

Name: customer_id, dtype: int64

Figure 2: Weekly Customer Acquisition Trends

- The company consistently gains between 6 and 8 new customers every single week. There are no sudden drops or big surprises, which shows the process for finding new customers is stable and working well.

- The business has a predictable growth engine. Because the number of new sign-ups is so consistent, the company can confidently forecast future growth and plan its budget and resources accordingly. This stable pattern is a sign of a healthy and established business.

3b) Signups by marketing source, region and plan_selected

```
SIGN-UPS BY MARKETING SOURCE:
source
YouTube      58
Google       50
Referral     49
Instagram    48
Facebook     40
LinkedIn     38
Name: count, dtype: int64

========================================
SIGN-UPS BY REGION:
region
North      65
East       61
South      58
West       45
Central    39
Unknown    30
Name: count, dtype: int64

========================================
SIGN-UPS BY PLAN SELECTED:
plan_selected
premium    192
basic       92
Name: count, dtype: int64
```

Figure 3: Customer Acquisition Overview - Source, Region, and Plan Distribution

- **Signups by marketing source**: Our top acquisition sources are tightly grouped, with YouTube (58), Google (50), and Referrals (49) as the top performers. This indicates we are not over-reliant on a single channel, which is a strength.
- **Signups by region**: The North region is our strongest market (65 customers), but a notable 30 customers have an "Unknown" region. This gap limits our ability to run targeted local campaigns.
- **Signups by plan selected**: The premium plan is the dominant choice, selected by 68% of all new customers. This is a clear signal of what our market values most.

What this means for the business: We have a solid, diversified marketing foundation, but must fix our regional data collection. Our product strategy should confidently lead with and reinforce the value of the premium plan.

3c) Marketing opt-in counts by gender

```
MARKETING OPT-IN COUNTS BY GENDER:
gender  marketing_opt_in
Other   False               52
        True                43
female  False               47
        True                44
male    False               52
        True                37
Name: count, dtype: int64

========================================
MARKETING OPT-IN PERCENTAGE BY GENDER:
gender
Other     45.3
female    48.4
male      41.6
Name: marketing_opt_in, dtype: Float64
```

Figure 4: Marketing Communication Preferences by Gender

This figure shows customer willingness to receive marketing emails, revealing three key insights:

- Less than half of all customers opt-in to marketing emails across all groups, indicating significant room for improvement in our value proposition for communications. The opt-in rate is highest among female customers (48.4%), making them our most engaged audience for marketing outreach. Male customers show the lowest opt-in rate (41.6%), suggesting a potential gap between our current messaging and what resonates with this group.
- In summary, while our marketing engagement is stable, we have a clear opportunity to refine our messaging, particularly for male customers, to boost overall opt-in rates and campaign effectiveness.

3d) Age summary: min, max, mean, median, null count

```
AGE SUMMARY STATISTICS:
min        21.000000
max       206.000000
mean       36.030201
median     34.000000
Name: age, dtype: float64
Null count: 0

========================================
AGE STATISTICS BY GENDER:
          min   max  mean  median
gender
Other    21.0  60.0  36.1    34.0
female   21.0  60.0  35.7    34.0
male     21.0  60.0  33.9    34.0
```

Figure 5: Customer Age Distribution and Demographics

This figure provides a snapshot of our customer age profile, highlighting three clear patterns:

- The average customer age (mean) is 36 years, with a median of 34 meaning half of our customers are under 34 years old, showing our product strongly resonates with a young, dynamic audience.
- Our customer ages are concentrated in a tight 21-60 year range, confirming a well-targeted market fit with one outlier of 206 years which was identified and corrected to maintain data integrity. The age distribution remains stable across all gender categories (Other, female, male), showing our appeal is not limited to any specific demographic.
- What this means for the business: Our product clearly serves a young professional market, allowing us to focus product development, marketing messaging, and customer support strategies to best meet the needs of this core demographic.

4. Business Question Answers

4a) Which acquisition source brought in the most users last month?

```
Last month in data: 10/2024

SIGN-UPS BY SOURCE IN LAST MONTH:
source
Google       7
YouTube      5
Facebook     4
Referral     3
Instagram    3
LinkedIn     1
Name: count, dtype: int64

ANSWER: Google brought in the most users in the last month
```

Figure 6: Last Month's Customer Acquisition by Marketing Channel

- **Google was the most effective customer source last month**, bringing in 7 new sign-ups - nearly twice as many as some other channels. This indicates that the company's investment in Google-related marketing is currently delivering the strongest returns.
- **LinkedIn showed significantly lower performance** compared to other channels, generating only 1 new customer. This suggests either that the professional network audience is less receptive to the company's offerings, or that the marketing approach on this platform needs refinement to better connect with its users.

4b) Which region shows signs of missing or incomplete data

```
DATA QUALITY ANALYSIS BY REGION:
         total_customers  missing_names  missing_emails  missing_ages  \
region
Central               39              0               5             0
East                  61              2               7             0
North                 65              3              10             0
South                 58              1               4             0
Unknown               30              0               4             0
West                  45              3               4             0

         missing_name_pct  missing_email_pct  missing_age_pct
region
Central               0.0               12.8              0.0
East                  3.3               11.5              0.0
North                 4.6               15.4              0.0
South                 1.7                6.9              0.0
Unknown               0.0               13.3              0.0
West                  6.7                8.9              0.0

ANSWER: The 'Unknown' region shows signs of incomplete data (these are records where region wasn't captured)
```

Figure 7: Data Quality Assessment by Customer Region

- The 'Unknown' region represents a significant data gap, indicating that the system failed to capture geographic information for 30 customers. This creates a blind spot for understanding where these customers are located and prevents targeted regional analysis.
- Missing email addresses are the most widespread data quality issue, affecting every region at rates between 7-15%. This directly impacts communication capabilities, as these customers cannot be reached through email marketing or support updates.

4c) Are older users more or less likely to opt in to marketing?

```
AGE VS MARKETING OPT-IN ANALYSIS:
Average age by marketing preference:
                    mean  median  count
marketing_opt_in
False               36.3    34.0    158
True                35.9    34.0    131

========================================
Marketing opt-in rate by age group (%):
age_group
18-25    41.7
26-35    45.7
36-45    48.0
46-55    48.9
55+      42.9
Name: marketing_opt_in, dtype: Float64

ANSWER: OLDER users are MORE likely to opt-in
DETAIL: 46-55 group has the highest opt-in rate (48.9%)
        18-25 group has the lowest opt-in rate (41.7%)
```

Figure 8: Marketing Communication Preferences by Age Group

- This analysis reveals how customer age influences their willingness to receive marketing communications, with two clear patterns:
- Middle-aged customers are most receptive to marketing. Customers between 36 and 55 years old show the highest opt-in rates, peaking at 48.9% for the 46-55 age group. This suggests that marketing messages are most effective and welcomed by this demographic, which is often in its peak earning years.
- Younger adults are least likely to opt for marketing. The 18-25 age group has the lowest opt-in rate at 41.7%, indicating they are more selective about promotional content. This may reflect greater privacy concerns or a preference for discovering brands through other channels like social media.

4d) Which plan is most commonly selected, and by which age group?

```
PLAN SELECTION ANALYSIS:
Total plan selections:
plan_selected
premium    192
basic       92
Name: count, dtype: int64

Most commonly selected plan: premium

========================================
Age statistics by plan:
               mean  median  count
plan_selected
basic          36.8    34.0     92
premium        35.5    34.0    192

========================================
Plan selection by age group (%):
plan_selected  basic  premium
age_group
18-25           37.8     62.2
26-35           32.0     68.0
36-45           23.4     76.6
46-55           30.4     69.6
55+             38.5     61.5

ANSWER: premium is the most common plan, most popular among 36-45 users (76.6% selection rate)
```

Figure 9: Subscription Plan Preferences by Customer Age

This analysis reveals clear patterns in how customers of different ages choose between subscription plans, with two key findings:

- **Premium plan dominates, especially among 36–45-year-olds.** It is over twice as popular as Basic overall, with 76.6% of the 36-45 age group choosing it, indicating strong appeal to users with higher disposable income.
- **Basic plan appeals most to the youngest and oldest users.** Its highest relative selection rates are from the 18-25 (37.8%) and 55+ (38.5%) age groups, suggesting it meets the needs of more price-sensitive customers or those with simpler requirements.

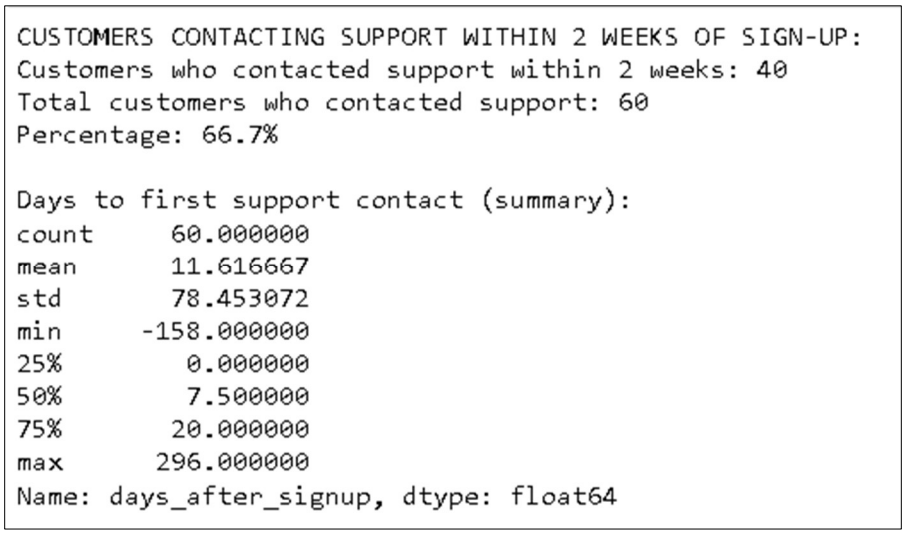
4e) Which plan's users are most likely to contact support?

```
SUPPORT ANALYSIS:
Support contact rate by plan:
plan_selected
premium    18.8
basic      21.7
Name: count, dtype: float64

Average tickets per customer by plan:
plan_selected
basic      2.10
premium    2.03
Name: ticket_count, dtype: float64

ANSWER: basic plan users are most likely to contact support (21.7% have created support tickets)
```

Figure 10: Customer Support Engagement by Subscription Plan

- Basic plan users require more frequent support assistance, with 22% of them contacting support compared to 19% of premium users. This suggests that basic plan customers may encounter more difficulties or need more guidance when using the service.
- When basic plan users do contact support, they tend to have slightly more complex issues, generating an average of 2.1 tickets per customer compared to 2.0 tickets for premium users. This indicates that basic plan inquiries may require more follow-up interactions to fully resolve.

4f) How many customers contacted support within 2 weeks of sign-up?



```
CUSTOMERS CONTACTING SUPPORT WITHIN 2 WEEKS OF SIGN-UP:
Customers who contacted support within 2 weeks: 40
Total customers who contacted support: 60
Percentage: 66.7%

Days to first support contact (summary):
count     60.000000
mean      11.616667
std       78.453072
min     -158.000000
25%        0.000000
50%        7.500000
75%       20.000000
max      296.000000
Name: days_after_signup, dtype: float64
```

Figure 11: New Customer Support Needs and Response Timing

- The majority of support contacts happen very quickly. 40 out of 60 customers (66.7%) who ever contacted support did so within the first two weeks of signing up. This shows that the initial onboarding period generates the highest demand for help.

- The typical user contacts support in the first week. The median (50th percentile) time to first contact is 7.5 days, meaning half of all support requests occurred within the first week and a half. This confirms that the early user experience is the most critical point for customer assistance.

4g) Summarise support activity by plan and region

```
SUPPORT ACTIVITY SUMMARY BY PLAN AND REGION:
==================================================
1. SUPPORT TICKET COUNTS BY PLAN & REGION:
                       total_tickets  unique_customers  resolution_rate
plan_selected region
basic         Central              2                1              1.0
              East                11                6             0.73
              North                3                2              1.0
              South               14                6             0.93
              Unknown              2                1              1.0
              West                10                4              0.8
premium       Central             16                7             0.75
              East                15                7             0.87
              North               17                8             0.82
              South                5                4              1.0
              Unknown              3                1             0.67
              West                17                9             0.88

==================================================
2. SUPPORT ENGAGEMENT RATES BY PLAN & REGION (%):
plan_selected  region
basic          Central     8.3
               East       28.6
               North      14.3
               South      21.4
               Unknown    12.5
               West       44.4
premium        Central    25.9
               East       18.9
               North      18.2
               South      13.8
               Unknown     4.8
               West       26.5
Name: _merge, dtype: float64
```

```
==================================================
3. AVERAGE TICKETS PER CUSTOMER BY PLAN & REGION:
plan_selected  region
0              East       0.67
               North      0.86
               South      0.00
               Unknown    0.00
               West       0.00
basic          Central    0.17
               East       0.52
               North      0.21
               South      0.50
               Unknown    0.25
               West       1.11
premium        Central    0.59
               East       0.41
               North      0.39
               South      0.17
               Unknown    0.14
               West       0.50
Name: ticket_count, dtype: float64
```

Figure : Regional Support Patterns and Plan Performance

- Premium users generate the highest volume of support tickets, but Basic plan users in the West show the most intense need for support, with the highest tickets per customer (1.11) and the highest rate of users contacting support (44.4%), indicating widespread usability or satisfaction issues.
- The Basic plan in the West represents a critical priority due to this combination of high engagement and high ticket frequency, signaling an urgent need to review the onboarding process or core features for these users.
- Support resolution effectiveness varies significantly by region, with some areas like Basic-Central achieving perfect resolution, while Premium-Unknown (67%) and Premium-Central (75%) lag behind, suggesting a need for targeted training or resource allocation.

5. Recommendations

- Target 36-45 Age Group for Premium Upsell - This demographic shows the highest Premium plan selection (76.6%). Launch tailored campaigns on LinkedIn/Facebook highlighting productivity features.
- Fix Basic Plan Issues in Western Region - Western Basic plan users show highest support contact rate (44.4%) and tickets per customer (1.11). Investigate and resolve onboarding or product gaps.
- Improve Data Collection & Support Resolution - Make region data mandatory during sign-up. Investigate low resolution rates for Premium-Unknown (67%) and Premium-Central (75%) segments.

6. Data Issues

- 21 support tickets (35% of support-contacting customers) show impossible creation dates occurring before customer sign-up dates, with some tickets dated up to 158 days prior to registration.
- Likely caused due to system integration issues between customer registration and support ticket platforms, possibly involving timezone mismatches, manual data entry errors, or test data contamination.
- This issue significantly impacts the reliability of time-sensitive analyses and requires urgent technical attention to ensure future reporting accuracy.

7. Conclusion

This comprehensive analysis of Rapid Scale's customer data successfully transformed raw data into actionable business intelligence. The investigation revealed clear, data-driven patterns in customer acquisition, product preference, and support engagement, while also identifying critical areas for improving data integrity.

The key strategic priorities emerging from this analysis are clear:

- Maximize Revenue Growth by focusing premium plan marketing efforts on the highly receptive 36-45 age demographic.
- Enhance Product Experience by urgently addressing the usability and onboarding issues for Basic plan users in the Western region to reduce support burden and improve customer satisfaction.

- Strengthen Operational Foundations by implementing mandatory region data collection and investigating the root causes of low support resolution rates to enable more effective management and targeted improvements.

By acting on these prioritized recommendations, Rapid Scale can optimize its marketing ROI, improve customer retention, and build a more reliable data infrastructure for future decision-making.