

Project Coversheet

Full Name	Alvin Siphosenkosi Moyo
Project Title (Example – Week1, Week2, Week3, Week 4)	Week1: Customer Sign-Up Behaviour & Data Quality Audit

Instructions:

Students must download this cover sheet, use it as the first page of their project, and then save the entire document as a PDF before submission.

Project Guidelines and Rules

1. Formatting and Submission

- Format: Use a readable font (e.g., Arial/Times New Roman), size 12, 1.5 line spacing.
- Title: Include Week and Title (Example - Week 1: Travel Ease Case Study.)
- File Format: Submit as PDF or Word file
- Page Limit: 4–5 pages, including the title and references.

2. Answer Requirements

- Word Count: Each answer should be within 100–150 words; Maximum 800–1,200 words.
- Clarity: Write concise, structured answers with key points.
- Tone: Use formal, professional language.

3. Content Rules

- Answer all questions thoroughly, referencing case study concepts.

- Use examples where possible (e.g., risk assessment techniques).
- Break complex answers into bullet points or lists.

4. Plagiarism Policy

- Submit original work; no copy-pasting.
- Cite external material in a consistent format (e.g., APA, MLA).

5. Evaluation Criteria

- Understanding: Clear grasp of business analysis principles.
- Application: Effective use of concepts like cost-benefit analysis and Agile/Waterfall.
- Clarity: Logical, well-structured responses.
- Creativity: Innovative problem-solving and examples.
- Completeness: Answer all questions within the word limit.

6. Deadlines and Late Submissions

- Deadline: Submit on time; trainees who fail to submit the project will miss the “Certificate of Excellence”

7. Additional Resources

- Refer to lecture notes and recommended readings.
- Contact the instructor or peers for clarifications before the deadline.

YOU CAN START YOUR PROJECT FROM HERE

1. Introduction

Rapid Scale requested a **data quality audit** and **customer sign-up behaviour analysis** to support the Monthly Business Review. The primary dataset (customer_signups.csv) was assessed for completeness, consistency, and readiness for segmentation, then used to analyse acquisition patterns, demographic trends, and marketing opt-in behaviour. An optional support dataset (support_tickets.csv) was reviewed to identify friction points and which customer groups contact support most frequently. The purpose of the report is to translate cleaned data into practical actions for Marketing, Product, and Customer Experience teams.

2. Data Cleaning Summary

The raw sign-up dataset contained **300 records** with multiple integrity issues: date and numeric fields stored as text, inconsistent categorical labels (plan, gender, opt-in), missing critical identifiers, and one duplicate customer ID. Records missing email were removed to preserve contactability and prevent unreliable user-level analysis, reducing the dataset to **264 valid customers**. Remaining missing values in region/source/gender were imputed as “Unknown” to retain analytical coverage. Helper fields (signup week/month and age band) were engineered to support cohort and demographic analysis. The cleaned dataset is now reliable for segmentation and business insight generation.

Metric	Before Cleaning	After Cleaning	Notes
Row count	300	264	34 email-missing rows removed + 1 missing customer_id row removed + 1 duplicate ID resolved
Missing email	34	0	Email required for contactability
Invalid dates	7	7 flagged as NaT	<3% of records affected
Duplicate IDs	1	0	Customer base now unique
Age missing	12 original + 7 invalid → 19	17	Remaining NaNs acceptable for segmentation
Plan categories	6 variants (Basic, BASIC, prem, UnknownPlan...)	3 unified categories	Prevents category fragmentation

Exhibit A: Before vs After Data Profile (rows, missingness, duplicates removed)

3. Key Findings & Trends

3.1 Acquisition & Plan Adoption

Acquisition is diversified across multiple sources, reducing dependency risk on any single channel. Plan selection is broadly balanced across Basic, Pro, and Premium, with a small residual “Unknown” category (<3%) reflecting data capture gaps rather than customer preference. Regionally, North and East show the strongest activity; however, the volume of “Unknown” region entries remains a material segmentation weakness and should be treated as a data capture risk.

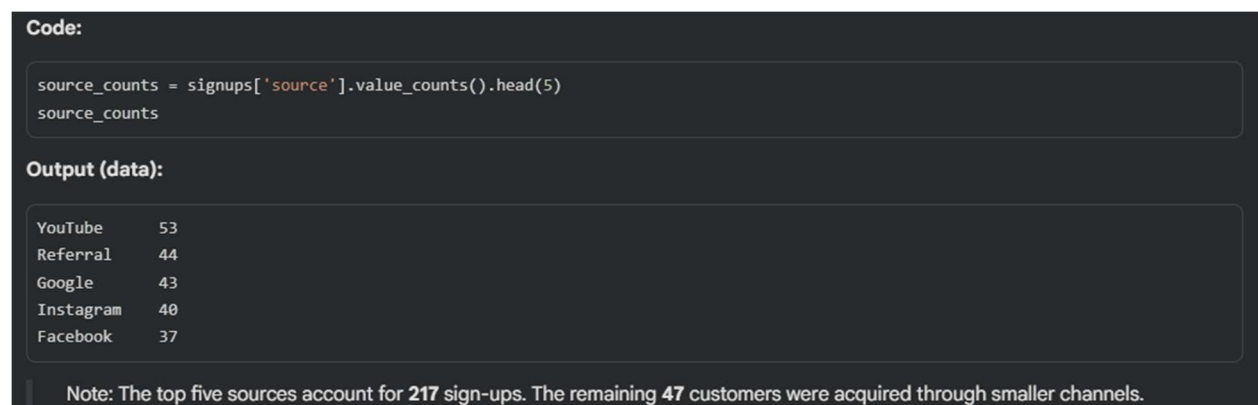


Figure 1: Sign-ups by Source (Top channels)

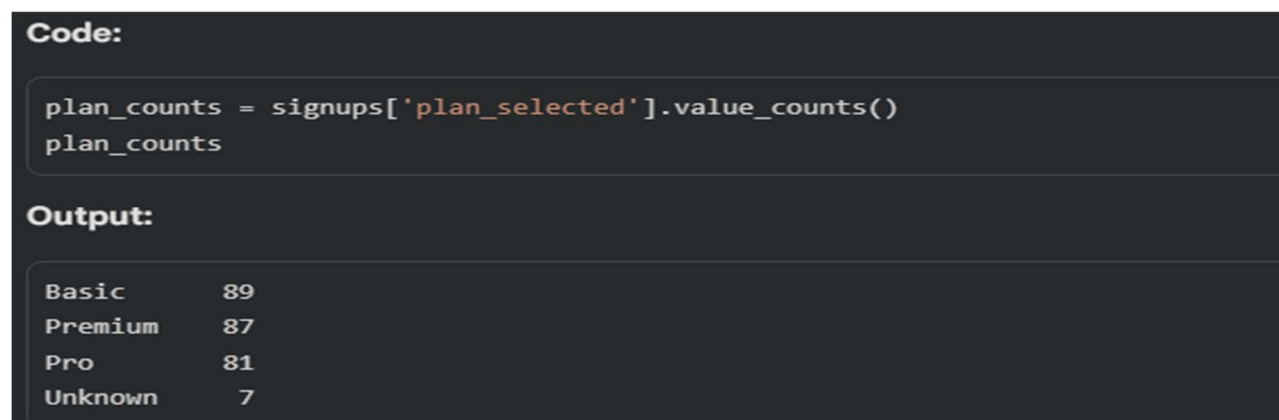


Figure 2: Overall Plan Adoption Distribution

3.2 Marketing Opt-In Behaviour

Opt-in rates vary by demographic group. Mid-aged cohorts show stronger consent patterns than younger users, and Female/Non-Binary customers show higher opt-in rates than Male users. This suggests marketing effectiveness will improve if campaigns are segmented by age band and demographic responsiveness, while younger users

may require alternatives to traditional email-first outreach (e.g., in-app prompts or creator-led messaging).

Code:

```
opt_in_age = signups.groupby('age_band')['marketing_opt_in'] \
    .value_counts(normalize=True) \
    .mul(100) \
    .unstack(fill_value=0)
```

opt_in_age

Output:

Age Band	No (%)	Yes (%)
18-25	64.3	35.7
26-35	54.5	45.5
36-45	52.0	48.0
46-55	51.1	48.9
56+	60.0	40.0

Note: Percentages exclude customers with missing age (n = 17).

Figure 3: Marketing Opt-In Rate by Age Band

Code:

```
opt_in_gender = signups.groupby('gender')['marketing_opt_in'] \
    .value_counts(normalize=True) \
    .mul(100) \
    .unstack(fill_value=0)
```

opt_in_gender

Output:

Gender	No (%)	Yes (%)
Female	51.1	48.9
Male	57.8	42.2
Non-Binary	51.3	48.7
Other	57.1	42.9

Figure 4: Marketing Opt-In Rate by Gender

3.3 Plan Selection by Age Band

Plan selection varies by age, following a gradual shift rather than a simple “Basic vs Premium” divide. The **18-25** cohort is Basic-heavy, reflecting price sensitivity, but **Pro (19) and Premium (17) adoption are closely aligned**, showing early openness to higher-tier plans. In the **26-35** group, **Premium (27) and Pro (26)** slightly exceed Basic (21), indicating growing perceived value. **Premium adoption peaks among 36–45** customers, while **46–55** shows a near-even distribution across plans before volumes taper in the 56+ group. Overall, Premium preference increases with age, but meaningful Pro and Premium uptake in younger cohorts supports targeted upgrade strategies rather than assuming strict price aversion.

Code:

```
plan_age = pd.crosstab(signups['age_band'], signups['plan_selected'])
plan_age
```

Output (formatted):

Age Band	Basic	Pro	Premium	Unknown	Total
18-25	27	19	17	2	65
26-35	21	26	27	2	76
36-45	13	11	20	1	45
46-55	15	16	15	1	47
56+	5	4	4	1	14

Note: Total = 247. The remaining **17 customers** are excluded from this view because they are missing Age data..

Table 5: Plan Selection by Age Band

Cross-tabulation showing the exact distribution of plan tiers across age cohorts.

3.4 Support Signals (optional dataset)

Support tickets are disproportionately concentrated in Basic and Pro, generating **more than twice the volume** of Premium support tickets. The two dominant friction points are **Login** and **Billing**, both high-impact lifecycle touchpoints that directly influence retention risk. The support pattern suggests usability and clarity gaps are more acute in lower tiers, while Premium appears to deliver a smoother experience.

Code:

```
merged_df = pd.merge(
    signups[['customer_id', 'plan_selected', 'region']],
    tickets,
    on='customer_id',
    how='inner'
)

support_activity = merged_df.groupby(['plan_selected', 'region'])['ticket_id'] \
    .count() \
    .unstack(fill_value=0)

support_activity['Total'] = support_activity.sum(axis=1)
support_activity.loc['Total'] = support_activity.sum()
support_activity
```

Output:

Plan	Central	East	North	South	West	Unknown	Total
Basic	2	11	6	14	7	2	42
Premium	3	0	6	2	8	0	19
Pro	8	14	11	3	4	3	43
Unknown	0	0	2	0	0	0	2
Total	13	26	24	19	19	5	106

Exhibit B: Support Tickets by Plan and Region

```
# Assuming the tickets DataFrame is loaded from the raw file
print("--- Top 5 Support Issue Types ---")
issue_counts = tickets['issue_type'].value_counts().head(5)
print(issue_counts.to_string())
```

```
--- Top 5 Support Issue Types ---
issue_type
Login Issue      29
Other            27
Billing          26
Technical Error  21
Account Setup    20
```

Exhibit C: Top Issue Types (Login, Billing)

4. Business Question Answers (Q1–Q5 only, each with evidence)

Q1. Which acquisition source brought in the most users last month?

Answer: YouTube is the top acquisition source, with Google/Instagram/Referral also contributing strongly. The channel mix is balanced, indicating resilient acquisition rather than reliance on a single platform.

Evidence: *Figure 1* (Sign-ups by Source) confirms the top channel and the relative distribution across the leading sources.

Implication: Maintain investment in the top-performing channel while protecting diversification through Google/Instagram/Referral to reduce platform risk.

Q2. Which region shows signs of missing or incomplete data?

Answer: The “Unknown” region category represents the main region completeness issue and limits geo-precision for marketing and support resourcing.

Evidence: *Exhibit A* (missing values profile) and the region distribution highlight the volume of missing region records imputed as “Unknown.”

Implication: Region should be made a required field upstream (validated dropdown) to prevent “Unknown” dilution in segmentation.

Q3. Are older users more or less likely to opt in to marketing?

Answer: Older and mid-aged cohorts are more likely to opt in than younger users, who show the lowest consent rates.

Evidence: *Figure 3* shows opt-in rates rising from younger to mid-aged groups.

Implication: Prioritise lifecycle marketing for higher-consent cohorts and design low-friction engagement for younger users (in-app/creator/community-led).

Q4. Which plan is most commonly selected, and by which age group?

Answer: Plan adoption is broadly balanced across tiers, with **Basic (89)** narrowly leading **Premium (87)** and **Pro (81)**. Premium uptake is strongest among mid-aged cohorts (36-45 and 46-55), while younger users (18-25) skew more toward Basic, with meaningful Pro and Premium adoption..

Evidence: *Figure 2* (Plan distribution) and *Table 5* (Plan selection by age band - cross-tabulation).

Implication: Marketing and pricing strategies should incorporate **age-banded messaging**, with targeted upgrade nudges focused on cohorts that already demonstrate higher Premium adoption.

Q5 (Optional). Which plan’s users are most likely to contact support?

Answer: Basic and Pro users are most likely to contact support; Premium users generate the fewest tickets despite similar plan adoption volume.

Evidence: *Exhibit B* shows ticket counts by plan/region; *Exhibit C* shows Login and Billing as dominant issue types.

Implication: Improve Basic/Pro onboarding and fix Login/Billing workflows to reduce friction and prevent churn risk.

5. Recommendations

- 1. Address high-friction touchpoints in Basic and Pro plans:** Prioritise improvements to **Login** and **Billing** workflows, which account for the majority of support tickets and are disproportionately raised by **Basic and Pro** customers. Streamlining authentication flows, clarifying billing information, and expanding self-service help content for these issues can materially reduce avoidable support volume and early customer frustration. These improvements should be treated as preventative retention measures rather than reactive support fixes.
- 2. Segment marketing by responsiveness and plan adoption:** Prioritise lifecycle and upgrade campaigns for **mid-aged cohorts (36-55)**, who demonstrate the highest marketing opt-in rates and stronger **Premium** adoption. In parallel, adopt **lighter-touch engagement strategies** (such as in-app prompts or creator-led content) for the **18-25** cohort, which shows lower responsiveness to traditional email-based marketing. This dual approach protects high-performing segments while allowing experimentation with harder-to-reach users.
- 3. Strengthen data capture for lifecycle and cohort analysis:** Improve upstream data collection by enforcing validated inputs for **region** and **source**, and by capturing **full sign-up timestamps** rather than dates alone. Enhanced timestamp precision would enable calculation of onboarding metrics such as **Time to First Support Ticket**, improving early-friction detection and enabling more targeted intervention strategies. Strengthening data capture at source will increase

confidence in future segmentation, retention modelling, and operational decision-making.

6. Data Issues & Risks

Despite successful cleaning and validation, the following data limitations remain and should be considered when interpreting results:

- **Missing or imputed geographic data:**

A non-trivial number of records required imputation to an "Unknown" region. This limits the precision of geo-targeted marketing and regional support analysis.

Mitigation: Enforce mandatory region selection via validated dropdowns during sign-up.

- **Dropped email-missing records may introduce bias:**

Removing customers without email addresses improves contactability but may exclude privacy-conscious or lower-engagement users, potentially skewing behavioural insights.

Mitigation: Monitor sign-up flows to understand why email capture fails and reduce drop-offs.

- **Timestamp precision limitation:**

The dataset records **sign-up dates only**, without time-of-day timestamps. This prevents calculation of **Time to First Support Ticket**, a key onboarding-health metric that would help distinguish early friction from later-stage issues.

Mitigation: Capture full timestamp (YYYY-MM-DD HH:MM:SS) at sign-up and in support logs to enable lifecycle latency analysis.
