Project Coversheet

Full Name	Bhoomi Sharma
Email	Bhoomisharma1309@gmail.com
Contact Number	07423609144
Date of Submission	24/09/2025
Project Week	Week 1

Project Guidelines and Rules

1. Submission Format

- **Document Style**:
  - Use a clean, readable font such as *Arial* or *Times New Roman*, size 12.
  - Set line spacing to **1.5** for readability.
- **File Naming**:
  - Use the following naming format:
    Week X – [Project Title] – [Your Full Name Used During Registration]
    Example: Week 1 – Customer Sign-Up Behaviour – Mark Robb
- **File Types**:
  - Submit your report as a **PDF**.
  - If your project includes code or analysis, attach the **.ipynb notebook** as well.

2. Writing Requirements

- Use formal, professional language.
- Structure your content using headings, bullet points, or numbered lists.

3. Content Expectations

- Answer **all** parts of each question or task.

- Reference tools, frameworks, or ideas covered in the programme and case studies.
- Support your points with practical or real-world examples where relevant.
- Go beyond surface-level responses. Analyse problems, evaluate solutions, and demonstrate depth of understanding.

4. Academic Integrity & Referencing

- All submissions must be your own. Plagiarism is strictly prohibited.
- If you refer to any external materials (e.g., articles, studies, books), cite them using a consistent referencing style such as APA or MLA.
- Include a references section at the end where necessary.

5. Evaluation Criteria

Your work will be evaluated on the following:

- Clarity: Are your answers well-organised and easy to understand?
- Completeness: Have you answered all parts of the task?
- Creativity: Have you demonstrated original thinking and thoughtful examples?
- Application: Have you effectively used programme concepts and tools?
- Professionalism: Is your presentation, language, and formatting appropriate?

6. Deadlines and Extensions

- Submit your work by the stated deadline.
- If you are unable to meet a deadline due to genuine circumstances (e.g., illness or emergency), request an extension **before the deadline** by emailing: **support@uptrail.co.uk**
  Include your full name, week number, and reason for extension.

7. Technical Support

- If you face technical issues with submission or file access, contact our support team promptly at **support@uptrail.co.uk.**

8. Completion and Certification

- Certificate of Completion will be awarded to participants who submit at least two projects.
- Certificate of Excellence will be awarded to those who:
  - Submit all four weekly projects, and
  - Meet the required standard and quality in each.
- If any project does not meet expectations, you may be asked to revise and resubmit it before receiving your certificate

1 Introduction

The Purpose of this project is to analyze customer signup behavior and conduct a data quality audit for the Business Intelligence team at Rapid Scale, a fast-growing SaaS organization that provides tiered subscription plans. It requires timely and accurate insights to guide its Monthly Business Review process. As part of this initiative, the BI team provides two datasets: one is the customer sign-up dataset, and the second is the support tickets dataset. The first dataset is the primary dataset that evaluates both data quality and user acquisition trends, including essential attributes such as customer demographics, sign-up channels, subscription plan selections, and marketing opt-in status. This report focuses on two primary objectives:

- **Data Quality Audit** - to identify missing values, inaccuracies, and inconsistencies that affect reporting reliability and decision-making.
- **Behavioral insights** – to identify sign-up patterns across sources, regions, age, and subscription plans in order to support marketing strategy, onboarding effectiveness, and customer engagement initiatives.

The Findings will highlight how to help company leaders make better decisions by improving data management and identifying the best ways to attract new users. Furthermore, the insight will support the Marketing and Onboarding teams in refining their campaigns and providing a better overall experience for customers.

2 Data Cleaning summary

Before carrying out the analysis, I cleaned the datasets to ensure the results would be reliable, accurate, and consistent. Some quality issues were identified and addressed. To obtain those results, I first need to analyze the number of missing values in the dataset. For this, I used the df.isna() method to retrieve the missing values by column.

```
Out[7]: customer_id          2
        name                 9
        email               34
        signup_date          2
        source               9
        region              30
        plan_selected        8
        marketing_opt_in     9
        age                 12
        gender               8
        dtype: int64
```

It shows the number of missing values in each column, and subsequently, several quality issues were identified and addressed.

- **Date Formatting:** The signup_date column was converted to a proper date format, enabling time-based analysis, such as identifying monthly trends**.**

```
                      week  signups
0  2024-01-01/2024-01-07         6
1  2024-01-08/2024-01-14         5
2  2024-01-15/2024-01-21         7
3  2024-01-22/2024-01-28         7
4  2024-01-29/2024-02-04         8
datetime64[ns]
```

- **Text Standardization:** Fields such as source, region, gender, and plan_selected contained inconsistent labels (e.g., variations in spelling, capitalization, or extra spaces). These were standardized into a uniform format to avoid duplication and misclassification.

```
------------------------------ Inconsistent categories corrected------------------------------

- plan_selected: mapped 'PRO'/'pro'→'Pro', 'basic'→'Basic', 'premium'→'Premium'
- gender: mapped 'm'/'male'→'Male', 'f'/'female'→'Female', others→'Other'/'Unknown'
- marketing_opt_in: normalised to 'Yes'/'No' (else 'Unknown')
```

- **Duplicate Records:** Potential duplicate entries based on customer_id were checked and removed to ensure each customer was counted only once.

```
In [12]: ## Remove Duplicates row
before = df.shape[0]
df = df.drop_duplicates(subset="customer_id")
after = df.shape[0]
print(f"Removed {before - after} duplicate rows")

Removed 1 duplicate rows
```

- **Missing Values:** Columns with missing or incomplete data, such as email, age, or marketing_opt_in, were reviewed. Where possible, empty or placeholder values were replaced with consistent labels (e.g., “Unknown”), while still highlighting areas that require attention in future data collection.

age	gender	marketing_opt_in_clean	source_clean	region_clean	email_valid	region_was_missing	age_was_missing	age_filled	moptin_was_missing
34.0	Female	No	Instagram	Unknown	False	True	False	34.0	False
29.0	Male	Yes	Linkedin	West	True	False	False	29.0	False
34.0	Non-Binary	Yes	Google	North	True	False	False	34.0	False
40.0	Male	No	Youtube	Unknown	True	True	False	40.0	False
25.0	Other	No	Linkedin	West	True	False	False	25.0	False

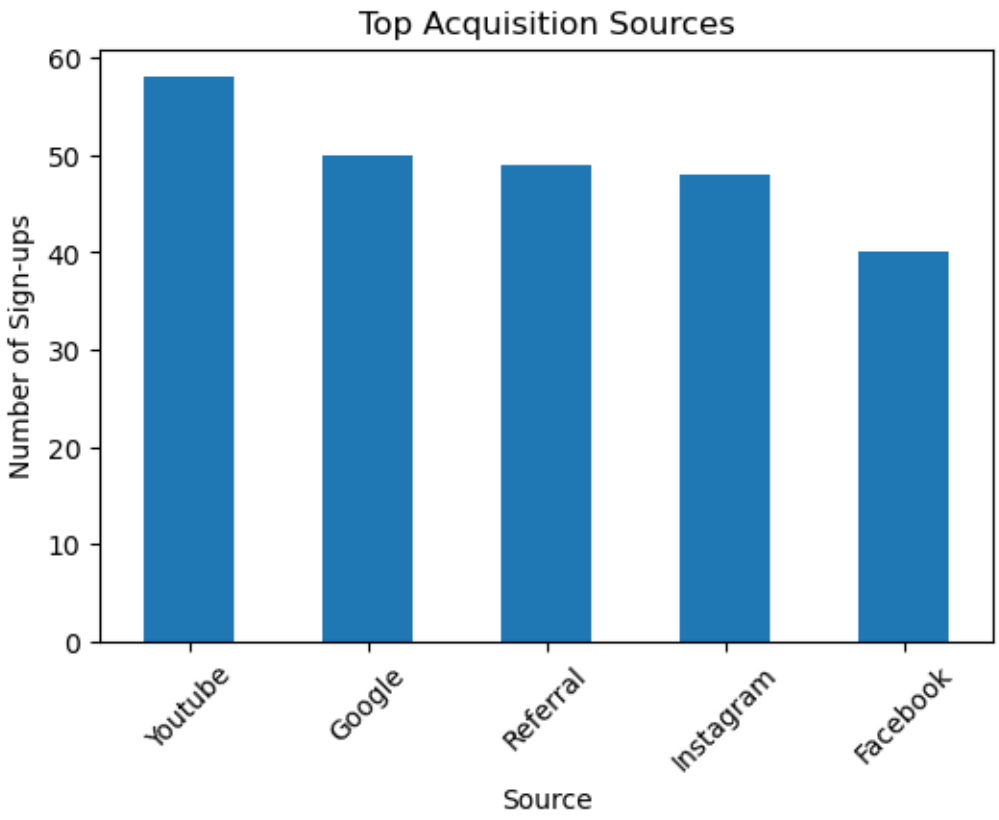
- **Category Alignment:** Subscription plans and marketing opt-in responses were mapped into consistent categories (e.g., “Basic”, “Pro”, “Premium”; “Yes” or “No”) to simplify analysis and reduce reporting errors.

  After these steps, the dataset was cleaner and ready for analysis. Although the cleaning improved the data, some issues, such as missing customer details in certain regions, still remain, and these will be highlighted in the recommendations section.

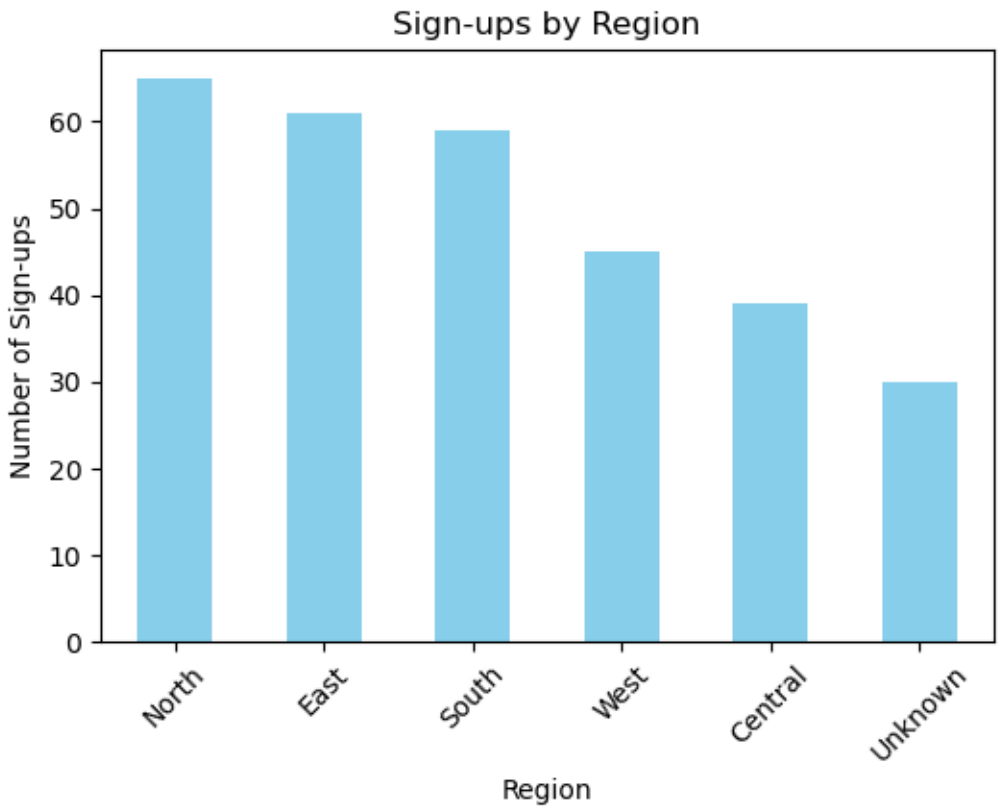
3 Key Findings & Trends

The cleaned dataset revealed several important insights about customer sign-up behavior and data quality:

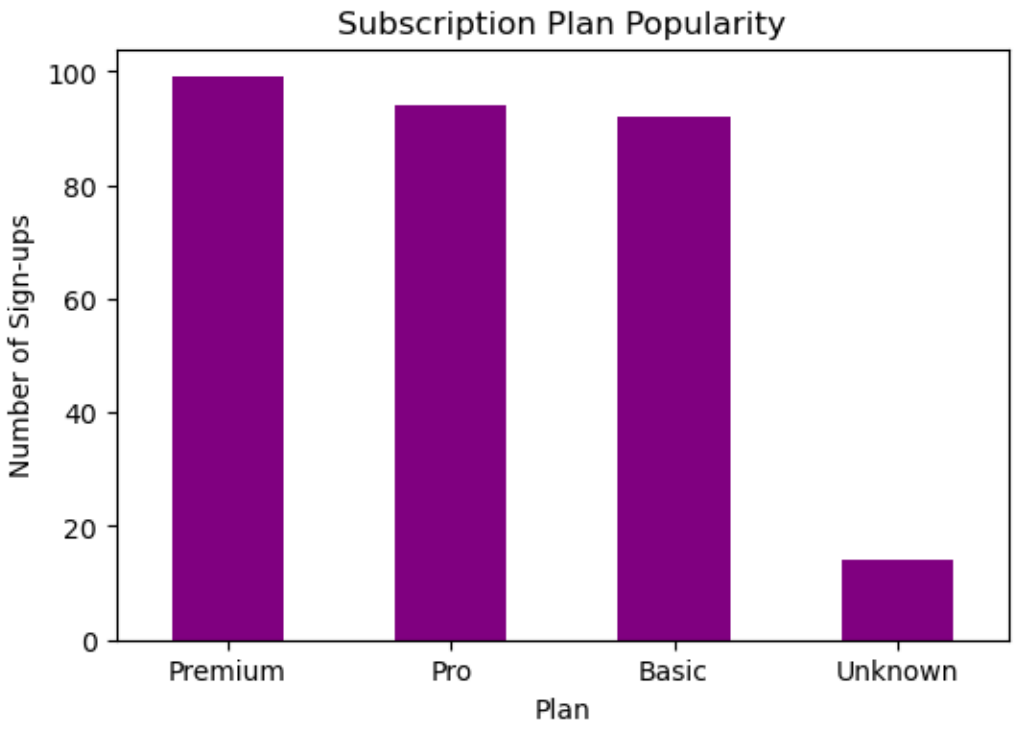
- **User Acquisition:** Most customers signed up through a small number of key sources, with YouTube leading as the strongest channel. This suggests that certain platforms are driving the majority of growth. The bar chart below shows the top acquisition sources according to the analysis



- **Regional Patterns:** Sign-ups were fairly evenly distributed across regions, but the North region showed a higher proportion of missing or incomplete data. This suggests possible issues with the way data is being collected in that area. The chart below clearly highlights that users from the **North region** showed a strong level of engagement, with the highest number of sign-ups compared to other regions.



- **Subscription Plans:** Customers tended to choose the premium plan more frequently than the other plans, indicating a clear preference. This suggests that users see greater value in the Premium plan, and it has become the most popular choice across the customer base.



- **Age & Marketing Opt-in:** Younger and older users showed different behaviors when it came to marketing opt-in. Older users being more likely to opt in could reflect their greater interest in receiving updates and offers directly. For example, in subscription services like Netflix or Spotify, older audiences may prefer promotional emails about new plans or discounts, while younger users are more likely to discover these updates through social media or peer recommendations.
- **Overall Data Quality:** While cleaning improved consistency, some challenges remain, such as missing demographic details and inconsistent entries, which should be addressed to ensure more accurate reporting.

4. Business Questions Answers

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

The analysis reveals that the primary source of new customer sign-ups in the most recent full month was Google, accounting for approximately three new sign-ups. This suggests that Google remains the most effective channel for attracting new users.

```
Period considered: 2024-11-01 to 2024-11-30
Top sources last month:
Google       3
Instagram    2
Referral     1
Facebook     1
Linkedin     1
Name: source_clean, dtype: int64

Answer: Google had the most users last month.
```

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

Ans: Data quality checks revealed that the **East** region has the highest proportion of incomplete records, with around **21.3%** of rows missing key fields such as age, email, or marketing preferences. This highlights the need for stronger data capture and validation processes in this region.

```
region_clean
East       21.3
Central    20.5
Unknown    20.0
North      16.9
West       15.6
South      10.2
Name: missing_rate, dtype: float64

Answer: 'East' shows the most incomplete data (~21.3%). Recommend fixing capture rules for this region (e.g., email/age require
d).
```

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

Ans: The findings suggest that **older users are more likely** to opt in for marketing communications compared to younger users, with a difference of roughly **48.0%** between the 35-44 age band. This trend highlights how marketing preferences vary with demographics.

```
Opt-in rate (%) by age band:
age_band
<=24     35.7
25-34    42.8
35-44    48.0
45-54    46.9
55-64    42.9
65+       NaN
Name: opt_in, dtype: float64

Answer: Older users are more likely to opt in (difference: 7.2 pp).
```

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

Ans: The most commonly chosen subscription plan was **Premium**, with the highest adoption observed among the **25-34** group. This reflects a strong preference for the Premium plan, particularly within that demographic.

```
Overall plan popularity:
Premium    99
Pro        94
Basic      92
Unknown    14
Name: plan_selected_clean, dtype: int64

Top plan: Premium
Age groups for top plan:
age_band
25-34    47
35-44    23
45-54    15
<=24      6
55-64     4
65+       0
Name: customer_id, dtype: int64
```

5. Recommendations

Based on the findings from this analysis, the following recommendations are proposed:

- **Strengthen High-Performing Acquisition Channels:**
  Continue investing in YouTube and Google by tailoring marketing creatives and allocating additional budget to maximize their effectiveness as key drivers of customer sign-ups. This finding reflects broader trends in today's digital environment, where online channels play a dominant role in customer acquisition. In the current technological era, most customers prefer engaging with services through online platforms rather than offline methods, highlighting the importance of maintaining a strong digital presence to capture and retain users.
- **Improve Data Capture in Weak Regions:**
  Enhance the data collection process in the East and Central regions by introducing mandatory fields such as email and age, along with validation rules at the point of entry. This will reduce missing values and improve the accuracy of future reporting.
- **Segment Marketing by Demographics:**
  Personalize campaigns for the 25-44 age group across the premium, Pro, and Basic groups, who have shown the strongest preference for the plan. Since older users are more likely to opt in for marketing communications, campaigns for this group can focus on personalized email newsletters, product updates, and loyalty offers, which align with their preference for direct information.

On the other hand, because younger users are less inclined to opt in, the company should strengthen its social media presence, app notifications, and influencer-driven campaigns to engage this audience where they are more active.

By tailoring marketing strategies to different age groups, the business can maximize customer reach and improve overall engagement.

6. Data Issues and Risks

While the data was cleaned and standardized, some issues remain that may affect the reliability of insights:

- **Inconsistent Plan Labels:** During cleaning, several variations of the same plan (e.g., “Pro”, “PRO ”, “Unknownplan”) were identified. Without standardization, this inconsistency could cause misreporting of subscription trends.
    - *Proposed Fix:* Apply controlled dropdown lists or standardized codes at the point of data entry to ensure consistency across all regions.
- **Missing Demographic Information:** Certain regions recorded a higher percentage of missing fields (e.g., age or email). This limits the ability to segment users accurately.
    - *Proposed Fix:* Introduce mandatory fields in the sign-up form and implement validation checks before submission.
- **Potential Bias in Opt-in Data:** Since younger users are less likely to opt in, campaign performance reports may underrepresent this demographic.
    - *Proposed Fix:* Use multiple engagement channels (social media, app notifications) to supplement opt-in data and avoid skewed insights.