Project Coversheet

Full Name	Mohammed Fahmidur Rahman
Email	20frahman@gmail.com
Contact Number	07429427096
Date of Submission	05/10/2025
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.

YOU CAN START YOUR PROJECT FROM HERE

Customer Sign-up Behaviour and Data Quality Analysis

This project involves analysing customer sign-up data for a SaaS company called Rapid Scale. As part of the Business Intelligence team, my goal is to audit the data quality, clean it, and then extract actionable insights about customer acquisition trends which will then be sent to the Marketing and Onboarding team to improve their campaigns and user engagement strategies. The dataset includes key customer details such as signup date, selected subscription plan, demographics (age and gender), marketing preferences, and more. Through this analysis, we aim to uncover patterns in how users sign up, identify data inconsistencies or gaps, and assess key findings and trends.

Data Cleaning Summary:

Before analysing the data, a thorough cleansing of the data was required to ensure accuracy and consistency.

Data Type Conversions

- The signup_date column was originally stored as a string which prevented time-based filtering or analysis, for this reason it was changed to proper datetime format using pd.to_datetime() to enable trend analysis over time.
- The age column was converted to numeric to support statistical summaries

Text Standardisation

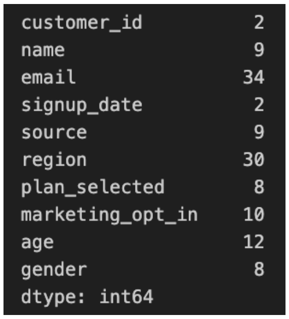
Several fields had inconsistent formatting within the dataset due to variations in text casing and spacing. For example, in the subscription plans column, entries like "PREMIUM" and "Premium" appeared as separate categories. These inconsistencies were resolved by stripping whitespace and applying consistent capitalisation across key categorical columns.

Inconsistent Category Corrections and Duplicate Removals

Specific incorrect or inconsistent values such as “Unknownplan” were corrected and standardised to “Unknown” to maintain consistency across categories. Additionally, there was one duplicate record that was found based on the customer_id column, this was likely a case where the user signed up more than once. This duplicate was removed, keeping only the first occurrence to ensure the customer is represented once in the dataset.

Missing Values

Several fields had missing values, as shown in the image below, which details the number of missing entries in each column.

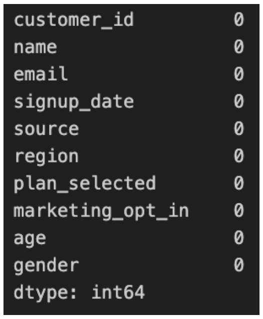


```
customer_id          2
name                 9
email               34
signup_date          2
source               9
region              30
plan_selected        8
marketing_opt_in    10
age                 12
gender               8
dtype: int64
```

To address these missing values, the following steps were taken to ensure data completeness and integrity:

- Filled missing categorical fields (source, region, plan_selected, gender, marketing_opt_in) with the placeholder value "Unknown" to retain these records while marking them as incomplete.
- Replaced missing email values with "unknown@unknown.com" to maintain valid email formatting without losing data.
- Converted age to numeric, coercing invalid entries to NaN, then filled missing ages with the median age of the dataset to preserve the overall distribution.
- Removed rows with missing or invalid signup_date, as this is a critical field for time-based analysis.

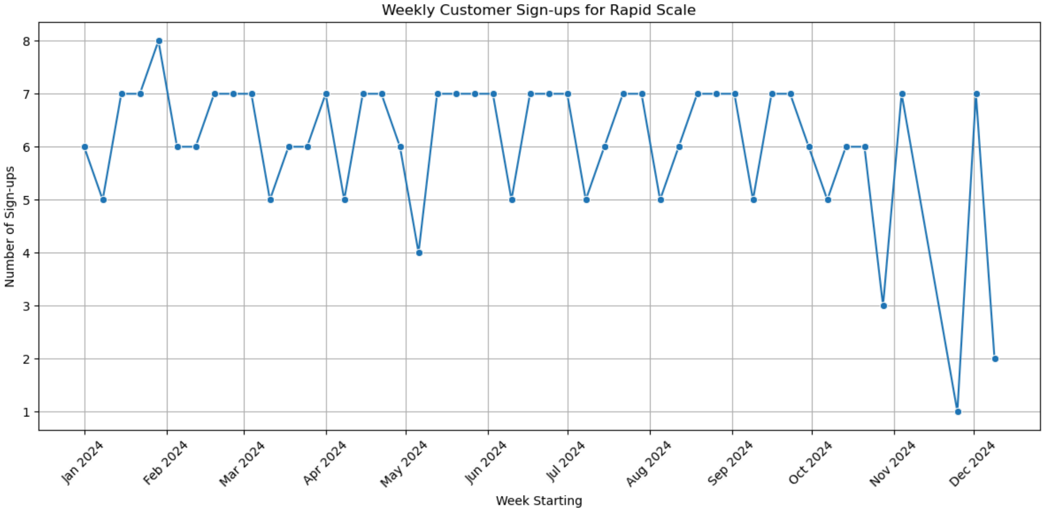
Following these cleaning steps, the dataset now contains no missing values, as demonstrated in the image below.



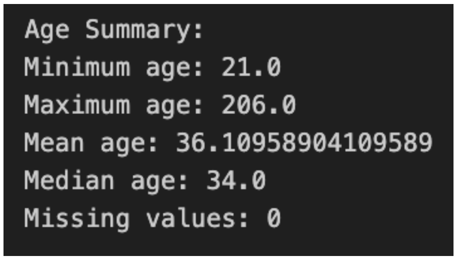
```
customer_id         0
name                0
email               0
signup_date         0
source              0
region              0
plan_selected       0
marketing_opt_in    0
age                 0
gender              0
dtype: int64
```

Key Findings and Trends:

1. **Stable Sign-up Rates:** Throughout the year, the user sign-up rate remained relatively consistent on a weekly basis of 5-7 sign-ups on average each week. However, there was a notable anomaly in mid-November, where sign-ups dropped sharply to just 1 user during the week of November 25 – December 1. This dip could indicate a temporary issue such as technical downtime or a paused marketing campaign.

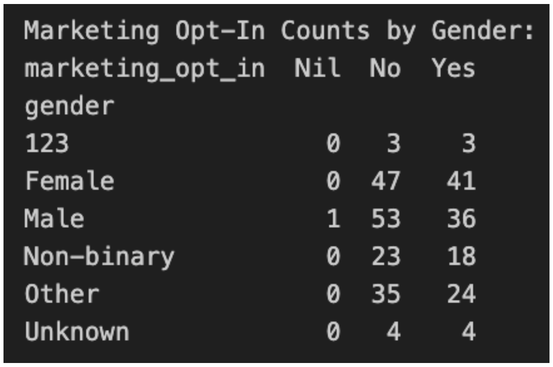


2. **User Age Distribution:** The average user age in the dataset is 36, suggesting that the platform is primarily attracting working-age adults. However, the recorded minimum and maximum ages are 21 and 206, respectively. While 21 is a reasonable minimum, the maximum age of 206 is clearly abnormal and indicates a data entry or validation issue. This type of anomaly could skew age-related analysis and should be addressed through stricter input validation at the point of data collection.



```
Age Summary:
Minimum age: 21.0
Maximum age: 206.0
Mean age: 36.10958904109589
Median age: 34.0
Missing values: 0
```

3. **Marketing Opt-In Patterns:** Marketing preferences were relatively balanced across genders. Female users were nearly evenly split between opting in and out, while males leaned slightly toward opting out. Non-binary and Other gender users showed higher opt-in rates proportionally. A data issue was noted with an invalid gender entry ("123").



```
Marketing Opt-In Counts by Gender:
marketing_opt_in  Nil  No  Yes
gender
123                 0   3    3
Female              0  47   41
Male                1  53   36
Non-binary          0  23   18
Other               0  35   24
Unknown             0   4    4
```

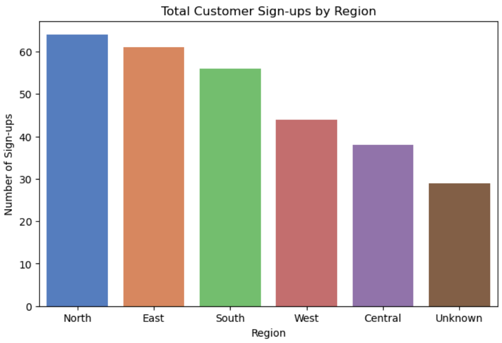
Business Question Answers:

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

Youtube was the highest customer acquisition source with Google, Referral, and Instagram following closely behind.

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

The dataset contains a significant number of users with the region value listed as "Unknown", indicating missing or unrecorded location data. This suggests that the region field itself is incomplete, rather than pointing to a specific region that is underrepresented.



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

The opt-in rate for users aged 18-64 remained fairly consistent with a percentage rate averaging around 45% throughout. Older users past the age of 64 are less likely to opt-in to marketing but users between 45-54 are likely to opt-in.

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

The most common plan overall was the Premium plan from the age group 35-44.

Recommendations:

- Target Marketing to Young Adults: The opt-in rate for young adults aged 18-24 is significantly lower than older adults aged between 25-64. Campaigns should be tailored more towards the needs and interests of this age group.
- Enhance Regional Data: The data within this dataset is limited to North, East, South, West, and Unknown. The campaign should include more regions as a significant number of users signed up with their region selected as Unknown, which likely indicates missing or incomplete data entry.

Data Issues or Risk:

An issue in this data is that there were several missing or inconsistent values in key fields such as email, plan_selected, and region. To fix this, strict validation rules must be implemented to ensure mandatory fields are completed accurately and consistently during data entry. Additionally, expanding the dropdown options and applying field constraints in future reporting processes will help reduce errors and ensure consistent formatting across the dataset.