Project Coversheet

Full Name	Maria Schiza
Email	mariaschiza005@gmail.com
Contact Number	07899940691
Date of Submission	13/07/2025
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

This report presents a data quality audit and exploratory analysis of the customer_signups.csv dataset provided by Rapid Scale, a SaaS company that operates on tiered subscription models. As a member of their Business Intelligence team, I conducted the following analysis using Python programming language to support the Monthly Business Review meeting. The results obtained aim to help the Marketing and Onboarding teams optimise user acquisition and engagement activities.

The customer_signups.csv dataset contains records of the following data fields:

- customer_id, which is a unique identifier for each customer
- name, email, age and gender of each customer
- signup_date, which is the date the customer registered for the service
- source, which is how the customer discovered the platform
- region, which is where the customer is based
- plan_selected, which is the subscription plan selected by the customer
- marketing_opt_in, which is whether the customer agreed to receive marketing communication

The objective of this project is to identify inconsistent, or duplicate records, to understand how users are discovering the platform, and which plans they choose, analyse demographic factors related to marketing opt-ins and understand data quality issues providing recommendations for future improvement.

2. Data Cleaning Summary

The dataset contained 300 records across 10 columns, including important user information. It is therefore necessary, to ensure data accuracy and reliability using the following cleaning steps before conducting meaningful analysis.

Data Types & Structural Validation

- The column signup_date contained values in mixed format and string types. These were converted using the dateutil.parser.parse() function, which can parse various datetime formats.
- The age column contained inconsistent numeric values such as the string ‘thirty’ instead of the number 30. These were converted using pd.to_numeric() with errors='coerce', which converted these values to numeric and flagged problematic entries as NaN.

Modifying these columns to the appropriate data type is important in ensuring compatibility with statistical analysis and visualisation tools.

Duplicate Removal

One duplicate record was removed using the. drop_duplicates() function on customer_id. Removing duplicate identifiers prevents bias and ensures an accurate analysis.

Standardisation of Categorical Text Values

In columns plan_selected and gender, inconsistent entries were found due to capitalisation differences or typos (e.g., "PRO", " prem"). These inconsistencies were corrected using .str.strip().str.title(). As shown in the Python output below, this is how these inconsistent values were corrected:

```
Inconsistent plan_selected values corrected:
 - basic → Basic
 - UnknownPlan → Unknownplan
 - PRO → Pro
 - prem → Premium
 - PREMIUM → Premium

Inconsistent gender values corrected:
 - male → Male
 - FEMALE → Female
```

In gender and source columns, any unexpected value (e.g., 123 and '??') has been converted to a null value. This approach aligns with best practices since any potential guess of their meaning would introduce inaccuracies.

Missing Values:

Missing values were identified across many columns using the **.**isnull().sum() command.

- The missing values in customer_id, email and signup_date fields have been removed since these columns contain critical information regarding trend analysis, user tracking, time-based analysis and communication.

- The missing values in the age column have been filled with the median value, 34. This reduces the influence of potential outliers and preserves reliability of results.
- Missing values in marketing_opt_in have been filled with the answer ‘No’ since there is clear absence of the answer ‘Yes’. This approach also prevents overestimation of opt-in rates and does not mislead the marketing team on the effectiveness of their acquisition channels.

The rest of the columns were left as null as they involve non-trivial context and guessing values would reinforce bias and incorrect behavioural assumptions.

3. Key Findings & Trends

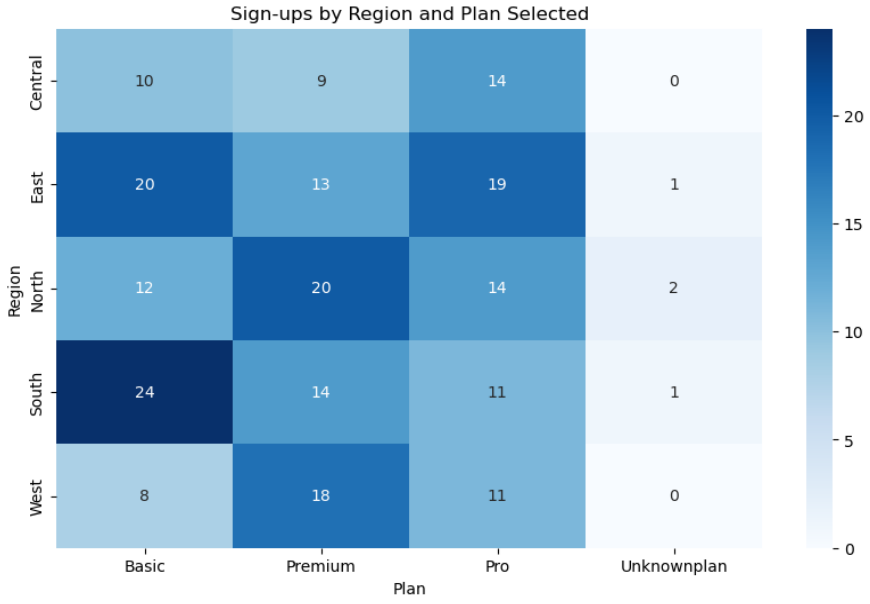
- Weekly Sign-Up Activity

  Weekly sign-ups have remained roughly consistent, averaging around 6 new customers per week. The highest activity was during the week of Jan 29 – Feb 4 reaching 8 signups, while the lowest during Mar 18 – Mar 24, Mar 25 – Mar 31, and Jul 15 – Jul 21, each with 5 signups.

  While growth appears stable, there is no sharp upward trend, suggesting potential strategic growth campaigns.

- User acquisition by source, region and plan

  Facebook and Google are the most acquired sources across all regions and plans, contributing to Premium and Pro tier signups, especially in the East and South. Sign-ups are geographically well-distributed, with slightly higher conversion toward Premium plans in Central and South regions, as shown in the heatmap.



This suggests that a successful approach to be taken is by increasing spend or content optimisation on Facebook and Google Ads targeting East and Central regions and Premium users.

- Marketing engagement by gender

  Across all gender groups, roughly 40–45% opted in to marketing communications, with female customers show a slightly higher engagement rate.

  The opt-in rate is moderate, suggesting that marketing consent capture could be improved—perhaps through incentives (e.g., discounts).

- Demographic profile
  The average age of customers is 36 years, with the median at 34, indicating that the target audience is of working professionals. The youngest user is 21, and the oldest is 60, with no nulls after data cleaning.

  Since there is a broad age range marketing messages should be tailored accordingly.

4. Business Question Answers

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

As shown from the table to the right, YouTube is the leading acquisition source, bringing in 6 new users, followed closely by Google and Referral, each with 5 users. This indicates that video content and search-based discovery are for the moment the most effective channels for user acquisition.

Most common acquisition source last month:

	source	count
0	YouTube	6
1	Google	5
2	Referral	5
3	LinkedIn	4
4	Facebook	3
5	Instagram	2

Given YouTube's performance, it may be valuable to further invest in or optimize our video campaigns and explore what content is driving conversions.

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

All regions from the dataset have shown missing information. However, the West region shows the highest proportion of incomplete data relative to its size with 11 missing fields across just 40 sign-ups equates to about 27.5% of entries containing at least one missing value, a percentage significantly higher than other regions.

This shows the possibility of data quality issues to how West region users are onboarded or how their data is recorded—perhaps a technical issue in the source form, system, or regional campaign.

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

Based on the data, users aged 25–35 have the highest marketing opt-in rate at 44.6%, followed closely by the 35–50 age group at 44.4%. On the contrary, the youngest group (18–25) and eldest group (50–65) show a slightly lower opt-in rate of 41.3% and 42.5% respectively.
Users in the middle age ranges (25–50) are slightly more likely to opt in to marketing compared to both younger and older users. While these are not significant differences, marketing efforts may see better engagement when targeted at users in the 25–50 age range.

IV. Which plan is most selected and by which age group?

The most selected plan across all users is the Premium plan, with 30 out of 85 users being aged 25-35. This insight can guide marketing strategies and plan feature enhancements targeting this age group.

```
Most commonly selected plan: Premium
Age groups selecting the most common plan:
          age plan_selected  count
1   (18, 25]       Premium     16
4   (25, 35]       Premium     30
8   (35, 50]       Premium     29
12  (50, 65]       Premium     10
```

5. Recommendations

1. Invest more of the marketing budget in social media, using targeted ads for the 25–40 age group. A similar approach is currently implemented by other SaaS companies like Spotify that tailor content based on age to increase sign-ups and keep customers longer. [2]
2. Standardise input fields by implementing dropdown lists or controlled vocabularies in all user-interactive forms. Add real-time input validation and mandatory fields for critical information (e.g., customer_id, email, signup_date) to reduce the rate of null entries.

3. Design a structured onboarding journey for Premium users, including welcome emails, and guided tutorials. Offer priority support for the first 30 days to reduce confusion.

6. Data Issues or Risks

- A frequent issue faced was inconsistent text entries such as varying capitalisations or misspellings in categorical columns which led to inaccurate aggregations during analysis. To address the issue there should be implemented dropdown menus or controlled vocabulary in the user sign-up interface.

- The dataset contains missing values in essential columns like customer_id and email, which are crucial for user tracking and communication. Dropping these rows results in data loss, reducing dataset completeness. Best way to address this is as mentioned before, by making these fields mandatory for users in providing valid information as well as real-time validation, such as email verifications.

7. References

All python functions referred in this pdf: Week 1 - Foundations of Data Analysis.pdf

[2] https://ads.spotify.com/en-GB/help-center/targeting-options/