Project Coversheet

Full Name	Omar Moustafa
Email	omoustafa@aucegypt.edu
Contact Number	+20 1023565867
Date of Submission	Thursday, July 24, 2025
Project 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

Project #1 Report

Customer Sign-Up Behaviour & Data Quality Audit

1. Introduction

This project thoroughly analyzed **Rapid Scale's customer sign-up dataset** in order to support the **Monthly Business Review** (MBR). The objectives were the following:

- Evaluate the **quality** of the **data**
- Understand and describe existing trends patterns about **user acquisition**
- Provide actionable and valuable insights for the **Marketing** and **Onboarding** teams

Two extensive and unclean datasets were provided:

- **customer_signups.csv** (primary dataset)
- **support_tickets.csv** (optional stretch, not included in this submission)

---

2. Data Cleaning Summary

Cleaning steps performed:

- **Converted** signup_date column to datetime format to conduct a time-based analysis

- **Standardised text entries**:

    - Corrected inconsistent casing in plan_selected (e.g. “PRO” → “Pro”) and gender columns.

    - Replaced unknown or invalid gender entries with “Unknown” to be as indicative as possible.

- **Removed duplicates**:

    - Dropped rows with duplicate customer_id values to ensure unique users and no user takes up multiple rows.

- **Handled missing values**:

    - Once again, missing values in the variable region were replaced with “Unknown”.

    - Replaced invalid age values, such as “unknown” and “thirty”, to name a couple, with NaN and imputed missing ages using the median (which was equal to 34).

    - Removed unrealistic age entries, such as age = 206, before computing summary statistics to make such summaries and analyses as realistic and reliable as possible.

- **Standardised marketing opt-in**:
    - Capitalized entries to keep a consistent theme and replaced invalid values such as “Nil” with “No”.

Some Relevant Outputs:

Data type inspection & missing value counts

(output of `customer_signups.info()` *and* `customer_signups.isnull().sum()`*)*

```
Customer Signups - Missing Values After Cleaning:
customer_id          1
name                 9
email               34
signup_date          6
source               9
region               0
plan_selected        8
marketing_opt_in    10
age                  0
gender               8
dtype: int64

Customer Signups - Information After Cleaning:
<class 'pandas.core.frame.DataFrame'>
Index: 299 entries, 0 to 299
Data columns (total 10 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   customer_id       298 non-null    object
 1   name              290 non-null    object
 2   email             265 non-null    object
 3   signup_date       293 non-null    datetime64[ns]
 4   source            290 non-null    object
 5   region            299 non-null    object
 6   plan_selected     291 non-null    object
 7   marketing_opt_in  289 non-null    object
 8   age               299 non-null    float64
 9   gender            291 non-null    object
dtypes: datetime64[ns](1), float64(1), object(8)
memory usage: 25.7+ KB
None
```

Final cleaned data preview *(output of `customer_signups.head()` after cleaning)*

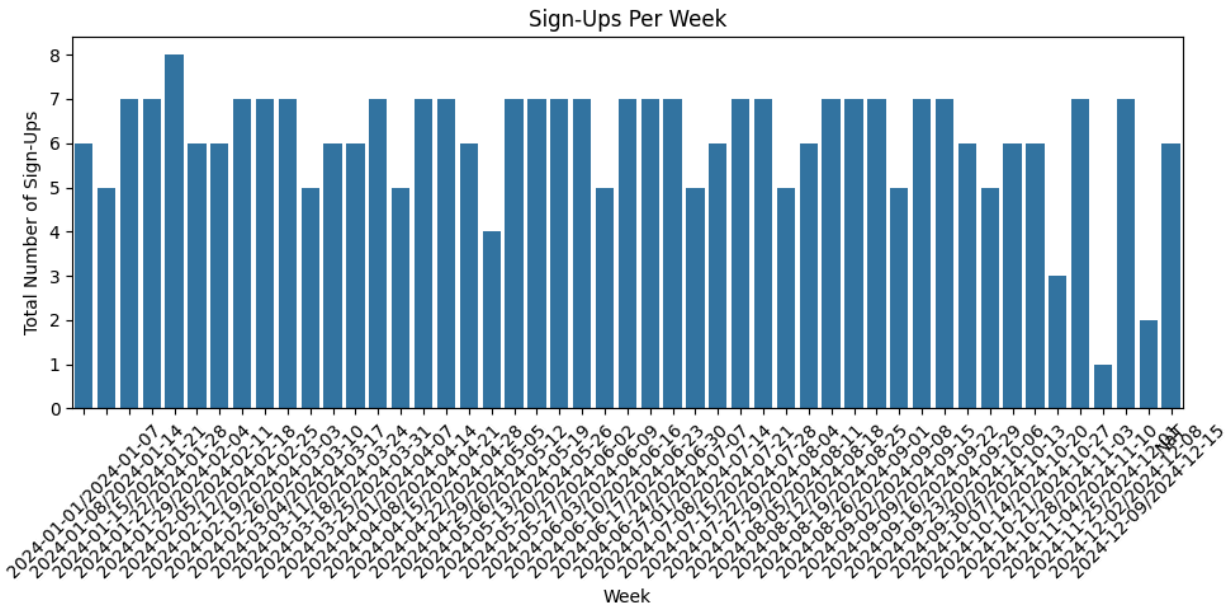
[65]:

	customer_id	name	email	signup_date	source	region	plan_selected	marketing_opt_in	age	gender	signup_week	age_group
0	CUST00000	Joshua Bryant	NaN	NaT	Instagram	Unknown	Basic	No	34.0	Female	NaT	31-40
1	CUST00001	Nicole Stewart	nicole1@example.com	2024-02-01	LinkedIn	West	Basic	Yes	29.0	Male	2024-01-29/2024-02-04	21-30
2	CUST00002	Rachel Allen	rachel2@example.com	2024-03-01	Google	North	Premium	Yes	34.0	Non-binary	2024-02-26/2024-03-03	31-40
3	CUST00003	Zachary Sanchez	zachary3@mailhub.org	2024-04-01	YouTube	Unknown	Pro	No	40.0	Male	2024-04-01/2024-04-07	41-50
4	CUST00004	NaN	matthew4@mailhub.org	2024-05-01	LinkedIn	West	Premium	No	25.0	Other	2024-04-29/2024-05-05	21-30

---

3. Key Findings & Trends

1. **The most common acquisition source** last month was **Google**. The runner-ups to Google were **Instagram** and **Referral**.



2. The **Premium** plan is the most selected plan, especially among **41-50 year olds**, while **21-30 year olds** are equally split on picking **Basic** and **Pro** plans.

3. **Marketing opt-in rates** are quite balanced across both genders, but **Non-binary users** tend to have slightly lower opt-in rates compared to Male and Female users.

---

4. Business Question Answers

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

✔ **Google** brought in the highest number of users last month.

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

✔ The **"Unknown" region** showed signs of missing/incomplete data.

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

✔ Analysis indicates there is **no clear trend** that suggests the tendencies of older users. With further and more extensive statistical testing that can validate some existing pattern or significance, Opt-in rates appear to remain similar across most age groups.

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

✔ The **Premium plan** is the most commonly selected plan, particularly by the age group of **41-50 year olds.** On the other hand, **21-30 year olds** tend to go for Basic and Pro plans, without a clear preference between the two.

---

5. Recommendations

1. **Target Premium upgrades** for **21-30 year olds**, as they form a large user base with room for upselling.

2. **Improve data collection** processes to reduce “Unknown” entries in region and gender fields for better segmentation.

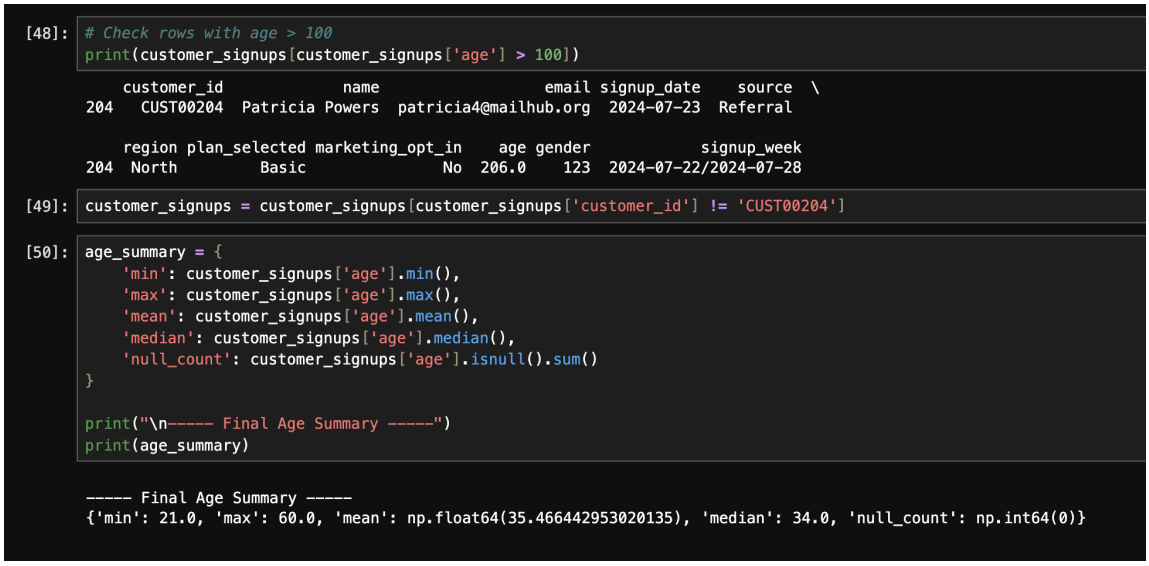
3. **Review onboarding flows** for Premium users aged **41-50**, ensuring features and benefits align with their usage to maintain retention.

---

6. Data Issues or Risks

Problem: Invalid and unrealistic age entries such as “thirty” as text or the maximum age being 206 years old, to name a couple.

Solution:
Implement **form validation rules at sign-up** to ensure age fields accept only numeric values within realistic human ranges (e.g. 18–100). This will reduce data cleaning efforts and necessities and improve the accuracy and reliability of reports.



```
[48]: # Check rows with age > 100
      print(customer_signups[customer_signups['age'] > 100])

          customer_id            name                 email signup_date    source  \
      204   CUST00204  Patricia Powers  patricia4@mailhub.org  2024-07-23  Referral

          region plan_selected marketing_opt_in    age gender           signup_week
      204  North         Basic              No  206.0    123  2024-07-22/2024-07-28

[49]: customer_signups = customer_signups[customer_signups['customer_id'] != 'CUST00204']

[50]: age_summary = {
          'min': customer_signups['age'].min(),
          'max': customer_signups['age'].max(),
          'mean': customer_signups['age'].mean(),
          'median': customer_signups['age'].median(),
          'null_count': customer_signups['age'].isnull().sum()
      }

      print("\n----- Final Age Summary -----")
      print(age_summary)

      ----- Final Age Summary -----
      {'min': 21.0, 'max': 60.0, 'mean': np.float64(35.466442953020135), 'median': 34.0, 'null_count': np.int64(0)}
```