Project Coversheet

Full Name	Andrea Aguirre
Email	andrea.vtag@gmail.com
Contact Number	083 893 5710
Date of Submission	05/08/2025
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

Introduction

As a new member of the Business Intelligence team at Rapid Scale, a fast-growing SaaS company, I was tasked with supporting the Monthly Business Review (MBR) by analysing customer sign-up data. This dataset captures essential information about new users, including subscription plans, sign-up channels, marketing preferences, and demographics such as gender and age. The insights generated will support the Marketing and Onboarding teams in optimising campaigns and engagement workflows.

The primary goal was to perform a **data quality audit** and identify trends in user acquisition. The Marketing and Onboarding teams wanted to understand:

1. *Data inaccuracies and incompleteness* that could cause issues
2. *Insight into how users are signing up* and which plans they're choosing
3. *Evaluating marketing opt-in behaviour* and user demographics

This report outlines how I assessed the dataset, cleaned the data, uncovered insights, and provided actionable recommendations using the analysis skills I have developed both in college and at Rapid Scale.

Data Cleaning Summary

I started by importing the dataset using Python's pandas library, ideal for data manipulation and analysis. After loading the CSV file, I examined its structure, data types, and missing values. The dataset had 300 records with missing values, notably in **email (34 missing)**, **region (30)**, and **age (12)**.

```python
# Identify missing values, data types, and column structure
import pandas as pd

# Loading dataset
fileName = 'customer_signups.csv'
data = pd.read_csv(fileName)
print(data)

print("\n\nData types")
# Checking types
print(data.dtypes)

print("\n\nMissing data")
# Detecting missing data
print(data.isnull().sum())

print("\n\nData structure")
# Checking structure
data.info()
```

```
Data types
customer_id         object
name                object
email               object
signup_date         object
source              object
region              object
plan_selected       object
marketing_opt_in    object
age                 object
gender              object
dtype: object


Missing data
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
```

```
Data structure
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 300 entries, 0 to 299
Data columns (total 10 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   customer_id       298 non-null    object
 1   name              291 non-null    object
 2   email             266 non-null    object
 3   signup_date       298 non-null    object
 4   source            291 non-null    object
 5   region            270 non-null    object
 6   plan_selected     292 non-null    object
 7   marketing_opt_in  290 non-null    object
 8   age               288 non-null    object
 9   gender            292 non-null    object
dtypes: object(10)
memory usage: 11.8+ KB
```

The *signup_date* column was converted to datetime format for consistency.

```python
# Convert signup_date to datetime
data['signup_date'] = pd.to_datetime(data['signup_date'])
# Output
print(data['signup_date'])
```

Next, I addressed the inconsistent text formatting by standardising key categorical columns. For example, ‘plan_selected’ had mixed casing and spelling variants, by using methods such as astype(), lower(), strip(), replace(), and capitalize().

```
# Standardise inconsistent text values (plan_selected, gender, etc.)
data['plan_selected'] = data['plan_selected'].astype(str).str.lower().str.strip().replace({
    'pro': 'Pro',
    'basic': 'Basic',
    'PREMIUM': 'Premium'
})
data['plan_selected'] = data['plan_selected'].str.capitalize()

# For gender column
data['gender'] = data['gender'].astype(str).str.lower().str.strip().replace({
    'MALE': 'Male',
    'FEMALE': 'Female',
    'male': 'Male',
    'female': 'Female'
})
data['gender'] = data['gender'].str.capitalize()

# For 'marketing_opt_in'
data['marketing_opt_in'] = data['marketing_opt_in'].astype(str).str.lower().str.strip().replace({
    'Nil': 'None',
})
data['marketing_opt_in'] = data['marketing_opt_in'].str.capitalize()
# Output
print(data['plan_selected'])
print(data['gender'])
print(data['marketing_opt_in'])
```

The same cleaning was applied to *gender* and *marketing_opt_in* fields to unify values like "MALE" to "Male" and "Nil" to "None".

While analysing the CSV file, I saw that there were duplicates. To remove duplicates based on the *customer_id*, I created a variable called *data_before*, which stores the number of rows in the dataset before removing the duplicates. The .shape[0] method returns the count of rows. The drop_duplicates() method removes any duplicate rows by adding subset='customer_id', and removes duplicates in the *customer_id* value that are repeated. Then I create another variable called *data_no_dupes*, which equals *data_before* subtracted by data.shape[0], and with a print statement, outputs the count, 1.

```
# Remove duplicate rows based on customer_id
data_before = data.shape[0]
data = data.drop_duplicates(subset='customer_id')
data_no_dupes = data_before- data.shape[0]
print(data_no_dupes)
```

Finally, missing values were handled thoughtfully, I used copy() to make a copy of the data to keep the original intact, the missing *region* was filled with "Unknown," by fillna(), age was converted to a numeric, by using to_numeric(), and missing values were replaced with the median age, median(), and records without emails were dropped, dropna(), due to the

importance of this field. I filled missing numeric values by interpolating between existing values, with interpolate(). Then printed each variable out.

```
# Handle missing values (e.g., region, email, age)
data_filled = data.copy()
data_filled['region'] = data_filled['region'].fillna('Unknown')
data_filled['age'] = pd.to_numeric(data_filled['age'])
data_filled['age'] = data_filled['age'].fillna(data_filled['age'].median())
data_filled = data_filled.dropna(subset=['email'])

data_dropped = data.copy()
data_dropped['age'] = pd.to_numeric(data_dropped['age'])
data_dropped = data_dropped.dropna()

data_interp = data.copy()
data_interp['age'] = pd.to_numeric(data_interp['age'])
data_interp = data_interp.interpolate()

print("Replaced:\n", data_filled)
print("Removed:\n", data_dropped)
print("Interpolate:\n", data_interp)
```

Replaced:

	customer_id	name	email	signup_date	\
1	CUST00001	Nicole Stewart	nicole1@example.com	2024-02-01	
2	CUST00002	Rachel Allen	rachel2@example.com	2024-03-01	
3	CUST00003	Zachary Sanchez	zachary3@mailhub.org	2024-04-01	
4	CUST00004	NaN	matthew4@mailhub.org	2024-05-01	
5	CUST00005	John Gonzales	john5@mailhub.org	2024-06-01	
..	...	...	...	...	
294	CUST00294	Mrs. Jessica Smith	mrs.94@example.com	2024-10-21	
295	CUST00295	Gary Smith	gary95@example.com	2024-10-22	
296	CUST00296	Anthony Roberts	anthony96@mailhub.org	2024-10-23	
298	CUST00298	Justin Mcintyre	justin98@mailhub.org	2024-10-25	
299	CUST00299	Mr. Bruce Bridges	mr.99@example.com	2024-10-26	

	source	region	plan_selected	marketing_opt_in	age	gender
1	LinkedIn	West	Basic	Yes	29.0	Male
2	Google	North	Premium	Yes	34.0	Non-binary
3	YouTube	Unknown	Pro	No	40.0	Male
4	LinkedIn	West	Premium	No	25.0	Other
5	Facebook	South	Premium	No	34.0	Other
..	...	...	...	...	...	...
294	Google	South	Pro	Yes	29.0	Other
295	Google	West	Premium	Yes	40.0	Nan
296	Google	Central	Basic	Yes	25.0	Female
298	YouTube	South	Premium	No	53.0	Male
299	NaN	North	Premium	Yes	29.0	Male

[265 rows x 10 columns]

Removed:

	customer_id	name	email	signup_date	\
1	CUST00001	Nicole Stewart	nicole1@example.com	2024-02-01	
2	CUST00002	Rachel Allen	rachel2@example.com	2024-03-01	
5	CUST00005	John Gonzales	john5@mailhub.org	2024-06-01	
6	CUST00006	Crystal Mason	crystal6@mailhub.org	2024-07-01	
7	CUST00007	Michael Bailey	michael7@mailhub.org	2024-08-01	
..	...	...	...	...	
293	CUST00293	Yvonne Harding	yvonne93@inboxmail.net	2024-10-20	
294	CUST00294	Mrs. Jessica Smith	mrs.94@example.com	2024-10-21	
295	CUST00295	Gary Smith	gary95@example.com	2024-10-22	
296	CUST00296	Anthony Roberts	anthony96@mailhub.org	2024-10-23	
298	CUST00298	Justin Mcintyre	justin98@mailhub.org	2024-10-25	

	source	region	plan_selected	marketing_opt_in	age	gender
1	LinkedIn	West	Basic	Yes	29.0	Male
2	Google	North	Premium	Yes	34.0	Non-binary
5	Facebook	South	Premium	No	34.0	Other
6	YouTube	North	Unknownplan	Yes	40.0	Male
7	YouTube	Central	Pro	Yes	60.0	Other
..	...	...	...	...	...	...
293	Google	Central	Pro	Yes	34.0	Male
294	Google	South	Pro	Yes	29.0	Other
295	Google	West	Premium	Yes	40.0	Nan
296	Google	Central	Basic	Yes	25.0	Female
298	YouTube	South	Premium	No	53.0	Male

[205 rows x 10 columns]

```
Interpolate:
     customer_id               name                  email signup_date  \
0     CUST00000       Joshua Bryant                    NaN         NaT
1     CUST00001      Nicole Stewart    nicole1@example.com  2024-02-01
2     CUST00002        Rachel Allen    rachel2@example.com  2024-03-01
3     CUST00003     Zachary Sanchez   zachary3@mailhub.org  2024-04-01
4     CUST00004                 NaN   matthew4@mailhub.org  2024-05-01
..          ...                 ...                    ...         ...
295   CUST00295          Gary Smith     gary95@example.com  2024-10-22
296   CUST00296     Anthony Roberts  anthony96@mailhub.org  2024-10-23
297   CUST00297  Timothy Mclaughlin                    NaN  2024-10-24
298   CUST00298     Justin Mcintyre   justin98@mailhub.org  2024-10-25
299   CUST00299   Mr. Bruce Bridges      mr.99@example.com  2024-10-26

        source   region plan_selected marketing_opt_in   age      gender
0    Instagram      NaN         Basic              No  34.0      Female
1     LinkedIn     West         Basic             Yes  29.0        Male
2       Google    North       Premium             Yes  34.0  Non-binary
3      YouTube      NaN           Pro              No  40.0        Male
4     LinkedIn     West       Premium              No  25.0       Other
..         ...      ...           ...             ...   ...         ...
295     Google     West       Premium             Yes  40.0         Nan
296     Google  Central         Basic             Yes  25.0      Female
297  Instagram     West         Basic             Yes  60.0         Nan
298    YouTube    South       Premium              No  53.0        Male
299        NaN    North       Premium             Yes  29.0        Male

[299 rows x 10 columns]
```

After filling in missing data, fixing errors, and removing inconsistencies, the dataset is now **clean and ready for analysis**.

Key Findings & Trends

During the data cleaning and analysis process, several significant trends and patterns emerged that help us better understand the customer dataset.

- Missing data was a notable issue across multiple columns. The **email** column had the highest percentage of missing values at **11.37%**, followed closely by **region** at **10.03%**. Other columns, such as age, marketing_opt_in, and gender, also contained some missing entries, ranging between 2.5% and 4%. The gaps were solved by using various methods, including filling missing regions with “Unknown”, replacing missing ages with the median, and removing records without email addresses due to their importance in identification and communication.

```
Missing Values (%):
 customer_id          0.334448
name                 3.010033
email               11.371237
signup_date          2.006689
source               3.010033
region              10.033445
plan_selected        2.675585
marketing_opt_in     3.344482
age                  4.013378
gender               2.675585
dtype: float64
```

- Inconsistencies in categorical data were corrected to improve data quality and analysis reliability. An example would be the *plan_selected* column has variations of ‘Pro’, like ‘pro’ and ‘PRO’, which were standardised to ‘Pro’. Similarly, the *gender* field had values like ‘MALE’ and ‘male’ that were unified under ‘Male’. The *marketing_opt_in* field was cleaned by mapping a variation of ‘Nil’ to a consistent format. These corrections ensured the categories were consistent, reducing errors in aggregation and reporting.

```
0       Basic
1       Basic
2     Premium
3         Pro
4     Premium
       ...
295   Premium
296     Basic
297     Basic
298   Premium
299   Premium
Name: plan_selected, Length: 300, dtype: object
0         Female
1           Male
2     Non-binary
3           Male
4          Other
         ...
295          Nan
296       Female
297          Nan
298         Male
299         Male
Name: gender, Length: 300, dtype: object
0      No
1     Yes
2     Yes
3      No
4      No
     ...
295   Yes
296   Yes
297   Yes
298    No
299   Yes
Name: marketing_opt_in, Length: 300, dtype: object
```

- The sign-up trends revealed a fairly stable pattern. Weekly sign-ups ranged mostly between 5 and 7 customers, experiencing small changes here and there, but nothing too extreme. This suggests steady customer acquisition without significant seasonal effects. The age distribution showed a mean age of approximately **36 years** and a median age of **34**.

```
Sign-ups per week by signup_date):
 signup_week
2024-01-01/2024-01-07    6
2024-01-08/2024-01-14    5
2024-01-15/2024-01-21    7
2024-01-22/2024-01-28    7
2024-01-29/2024-02-04    8
2024-02-05/2024-02-11    6
2024-02-12/2024-02-18    6
2024-02-19/2024-02-25    7
2024-02-26/2024-03-03    7
2024-03-04/2024-03-10    7
2024-03-11/2024-03-17    5
2024-03-18/2024-03-24    6
2024-03-25/2024-03-31    6
2024-04-01/2024-04-07    7
2024-04-08/2024-04-14    5
2024-04-15/2024-04-21    7
2024-04-22/2024-04-28    7
2024-04-29/2024-05-05    6
2024-05-06/2024-05-12    4
2024-05-13/2024-05-19    7
2024-05-20/2024-05-26    7
2024-05-27/2024-06-02    7
2024-06-03/2024-06-09    7
2024-06-10/2024-06-16    5
```

```
Age Summary:
 min        21.000
max       206.000
mean       36.175
median     34.000
Name: age, dtype: float64
Null count in age: 19
```

Overall, these results point to opportunities to enhance how we collect data, while also showing consistent customer activity throughout the year, providing a solid base for deeper analysis.

Business Question Answers

1. Based on the signup data grouped by the *source* column, the last month, October 2024, shows **Google** as the leading acquisition source, bringing in the highest number of users

compared to others like YouTube or LinkedIn. This comes as no surprise since Google is widely used and easily accessible to a vast audience, making it a common choice for users when discovering new services or products.

2. The *region* column has a significant amount of missing values (around 10%), indicating incomplete data, especially for regions such as Central and Unknown, where many entries have missing or unclear region info. This can happen because some customers do not provide their location, for privacy concerns.
3. Older users tend to be less likely to opt in to marketing. The marketing opt-in counts show younger age groups have higher "Yes" responses, while older users show a more balanced or lower opt-in rate. This suggests that age can influence user behaviour, particularly when it comes to trust, preferences, and digital engagement.
4. The **Premium plan** is the most commonly selected across the dataset. It is shown to be popular among users in their 30s and 40s, which is close to the average age range of the data set. This is likely because the 30-40-year-old working class have more financial stability and tends to seek higher-value services, which makes the Premium plan more appealing to them.

Recommendations

1. Improving Data Collection Process

   A significant number of entries had missing or inconsistent information, especially in the region, email, and age fields. To enhance data quality, implement mandatory field checks during user sign-up and introduce validation rules to reduce missing or incorrectly formatted data.

   I noticed that the *plan_selected* column includes an entry labelled ‘UnknownPlan.’ To improve consistency, it should be limited to three valid options plus a ‘None’ category for customers without a selected plan.

   Inconsistent age entries, like the word "thirty" instead of the number 30, and ‘Unknown’, were found. To maintain clean and analyzable data, age inputs should be restricted to numeric values only.

2. Target Marketing Toward Engaged Age Groups

   Users in their 30s and 40s were more likely to choose higher-tier plans and opt in to marketing communications. Focusing future campaigns and promotions on this demographic makes sense, as they demonstrate high levels of engagement and are likely to convert into paying customers.

3. Optimise Acquisition via Top-Performing Channels

   Google consistently attracts the most new users, particularly in recent months. To maximise this strength, consider increasing investment in Google Ads and enhancing SEO efforts to attract higher-quality traffic.

Data Issues or Risks

A notable data quality issue was the unrealistic age value of **206**, suggesting user entry or data integrity errors. Implementing age limits (**e.g., 18 to 100**) during data entry will prevent such anomalies. For future reporting, automated data quality checks could flag or filter outliers before analysis to ensure more accurate and reliable insights.