Project Coversheet

Full Name	Rory Scott
Email	Rdscott40ksx@hotmail.co.uk
Contact Number	+44 7590235745
Date of Submission	11/07/2025
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.

YOU CAN START YOUR PROJECT FROM HERE

Uptrail Internship Program Week 1 Project

Customer Sign-Up Behaviour & Data Quality Audit

By Rory Scott

Introduction

As a new analyst at Rapid Scale, I was asked to review recent customer sign-up data. Rapid Scale is a growing SaaS company with tiered subscription plans, and this analysis will support Marketing and Onboarding teams.

The goals of this project are:

1. To carry out a **data quality audit** by identifying missing, inconsistent, or duplicate entries.
2. To uncover **user acquisition trends**, including how users are signing up, which plans they're choosing, and how different age groups or regions are engaging.

The work was done using **Python**, **Pandas**, **NumPy**, and **Jupyter Notebook** to clean and analyse the data. The end goal is to provide clear, useful insights backed by accurate data.

Section 1: Data Loading & Cleaning

📌 Introduction

In this section, we load the customer sign-up data and check for issues. We fix data types, clean up categories, remove duplicates, and handle missing values to make the dataset ready for analysis.

Part 1: Identifying Data Frame Structure

To start the audit, we used df.info() to check the structure of the dataset. This gave us:

1. **300 rows** and **10 columns** of customer data.

2. All columns were stored as **object type**, meaning they are treated as text. Some should be converted (e.g. signup_date to datetime, age to numbers).

3. **Missing values** were found in every column.

```
<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: 23.6+ KB
```

This step helped us understand what needed cleaning before analysis.

Part 2: Identifying Unique Column Values

Before proceeding with cleaning, it's important to understand what kinds of values exist in each column.

The output to the right shows the raw unique values present in each key column before any cleaning was applied.

```
📊 Raw unique values in 'plan_selected':
['basic' 'PREMIUM' 'Pro' 'Premium' 'UnknownPlan' 'PRO' 'Basic' nan 'prem']

📊 Raw unique values in 'gender':
['Female' 'Male' 'Non-Binary' 'Other' 'male' 'FEMALE' nan '123']

📊 Raw unique values in 'age':
['34' '29' '40' '25' '60' '47' '53' '21' nan 'unknown' 'thirty' '206']

📊 Raw unique values in 'marketing_opt_in':
['No' 'Yes' nan 'Nil']

📊 Raw unique values in 'region':
[nan 'West' 'North' 'South' 'Central' 'East']

📊 Raw unique values in 'source':
['Instagram' 'LinkedIn' 'Google' 'YouTube' 'Facebook' 'Referral' nan '??']

📊 Raw unique values in 'signup_date':
[nan '02-01-24' '03-01-24' '04-01-24' '05-01-24' '06-01-24' '07-01-24'
 '08-01-24' '09-01-24' '10-01-24' '11-01-24' '12-01-24' '13-01-24'
 '14-01-24' '15-01-24' '16-01-24' '17-01-24' '18-01-24' '19-01-24'
 '20-01-24' '21-01-24' '22-01-24' '23-01-24' '24-01-24' '25-01-24'
 '26-01-24' '27-01-24' '28-01-24' '29-01-24' '30-01-24' '31-01-24'
 '01-02-24' '02-02-24' '03-02-24' '04-02-24' '05-02-24' '06-02-24'
 '07-02-24' '08-02-24' '09-02-24' '10-02-24' '11-02-24' '12-02-24'
 '13-02-24' '14-02-24' '15-02-24' '16-02-24' '17-02-24' '18-02-24'
 '19-02-24' '20-02-24' '21-02-24' '22-02-24' '23-02-24' '24-02-24'
 '25-02-24' '26-02-24' '27-02-24' '28-02-24' '29-02-24' '01-03-24'
 '02-03-24' '03-03-24' '04-03-24' '05-03-24' '06-03-24' '07-03-24'
 '08-03-24' '09-03-24' '10-03-24' '11-03-24' '12-03-24' '13-03-24'
 '14-03-24' '15-03-24' '16-03-24' '17-03-24' '18-03-24' '19-03-24'
 '20-03-24' '22-03-24' '23-03-24' '24-03-24' '25-03-24' '26-03-24'
 '2nd february 2024' '28-03-24' '29-03-24' '30-03-24' '31-03-24'
 '01-04-24' '02-04-24' '03-04-24' '04-04-24' '05-04-24' '06-04-24'
```

These insights inform the cleaning approach for each column in the next steps.

📌 Part 3: Handling Datatypes

`Before analysis, it's important to make sure each column has the right data type. For example, signup_date was changed to datetime for time-based grouping, and age was made numeric for calculations. This helps avoid errors and supports accurate analysis and visualisation.

```
<class 'pandas.core.frame.DataFrame'>
Index: 298 entries, 0 to 299
Data columns (total 10 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   customer_id       298 non-null    object
 1   name              289 non-null    object
 2   email             264 non-null    object
 3   signup_date       292 non-null    datetime64[ns]
 4   source            283 non-null    object
 5   region            268 non-null    object
 6   plan_selected     284 non-null    object
 7   marketing_opt_in  298 non-null    object
 8   age               298 non-null    Int64
 9   gender            284 non-null    object
dtypes: Int64(1), datetime64[ns](1), object(8)
memory usage: 25.9+ KB
```

📌 Part 4: Handling Inconsistent Column Values

The outputs below display the transformation of raw categorical values for each column into standardised formats. This process ensures consistency in categorical data and helps maintain data quality for accurate analysis.

Plan Selected & Gender: Inconsistent text formats (e.g. casing differences like 'PRO', 'pro') were standardised into clean categories (e.g. 'Pro', 'Female'). Final totals and unmapped (null) values are clearly reported.

```
🔄 Original → Cleaned Breakdown for 'plan_selected':
'Basic' → 'Basic': 46
'PREMIUM' → 'Premium': 42
'PRO' → 'Pro': 41
'Premium' → 'Premium': 57
'Pro' → 'Pro': 53
'basic' → 'Basic': 46
'prem' → 'Premium': 1

✅ Final Totals by Cleaned Category:
'Basic': 92
'Premium': 100
'Pro': 94

⚠️ Null values in 'plan_selected': 14
```

```
🔄 Original → Cleaned Breakdown for 'gender':
'FEMALE' → 'Female': 52
'Female' → 'Female': 41
'Male' → 'Male': 44
'Non-Binary' → 'Non-Binary': 42
'Other' → 'Other': 59
'male' → 'Male': 48

✅ Final Totals by Cleaned Category:
'Female': 93
'Male': 92
'Non-Binary': 42
'Other': 59

⚠️ Null values in 'gender': 14
```

Marketing Opt-In & Region: Irregular entries like 'Nil' or inconsistent capitalisation were cleaned to standard values. Cleaned category counts and remaining nulls are shown.

```
🔁 Original → Cleaned Breakdown for 'marketing_opt_in':
'Nil' → 'No': 1
'No' → 'No': 156
'Yes' → 'Yes': 133

✅ Final Totals by Cleaned Category:
'No': 157
'Yes': 133

⚠ Null values in 'marketing_opt_in': 10
```

```
🔁 Original → Cleaned Breakdown for 'region':
'Central' → 'Central': 39
'East' → 'East': 61
'North' → 'North': 65
'South' → 'South': 59
'West' → 'West': 46

✅ Final Totals by Cleaned Category:
'Central': 39
'East': 61
'North': 65
'South': 59
'West': 46

⚠ Null values in 'region': 30
```

Age: Non-numeric values (e.g. 'unknown', 'thirty') were detected and set as null. Valid numeric entries were retained.

```
Cleaning numeric column: 'age'
→ Non-null values: 280
→ Null values in 'age': 20

⚠ Invalid (non-numeric) original values that caused NaN:
'unknown': 6 time(s)
'thirty': 1 time(s)
```

Signup Date: The signup_date column was successfully converted to datetime for most entries. One unparsable value and a few nulls were identified and handled.

```
Cleaning date column: 'signup_date'
‣ Parsed 294 values
‣ Null values: 6

⚠ Unparseable values (1): ['not a date']
```

Source: Invalid values ('Nan', '??') were removed, leaving six clear source categories. Some entries remained null.

```
📝 Cleaning text column: 'source'
→ Excluded 2 invalid values: ['Nan' '??']
→ Unique values: 6
→ Null values: 15
→ Sample values: ['Instagram' 'Linkedin' 'Google' 'Youtube' 'Facebook']
```

These outputs ensure both categorical and numeric fields are clean and ready for accurate analysis.

📌 Part 5: Identifying Missing Values

Identifying missing values at this stage helps prioritise which fields need imputation, exclusion, or further investigation in the cleaning process.

	Column Name	Missing Count (Before)	% Missing (Before)
2	email	34	11.33
5	region	30	10.00
8	age	12	4.00
7	marketing_opt_in	10	3.33
4	source	9	3.00
1	name	9	3.00
9	gender	8	2.67
6	plan_selected	8	2.67
0	customer_id	2	0.67
3	signup_date	2	0.67

The table above highlights the number and percentage of missing values in each column before cleaning.

📌 Part 6: Handling Missing Values and Unique Identifier Duplicates

To address missing data, each column was reviewed and handled based on its role in the analysis.

```
❌ Removed 2 row(s) with missing customer_id.
🔁 Removed 0 duplicate row(s) based on customer_id.

📊 Missing Value Handling Summary:
Column               Handled Count   Method
-----------------------------------------------------
customer_id          2               Rows removed
marketing_opt_in     10              Filled with "No"
age                  20              Filled with median
```

The table shows how missing values were handled for key columns:

- **Customer ID:** 2 missing records were removed, as each user needs a unique ID.
- **Marketing Opt-In:** 10 missing entries were assumed to mean "No".
- **Age:** 20 missing values were filled with the median to avoid skewing the data.
- **Duplicate/missing** customer IDs were also removed to ensure each user is unique.

This method kept data loss low while maintaining consistency.

Section 2: Data Quality Overview

📌 Introduction

This section reviews the dataset's quality after cleaning, highlighting missing values, removed duplicates, and corrected category issues to ensure reliable analysis.

📌 Part 1: Missing Values Overview (After Cleaning)

The table below presents the number and percentage of missing values remaining in each column after applying tailored cleaning methods.

	Column Name	Post-handling Missing Count	% Missing
0	customer_id	0	0.00
1	name	9	3.02
2	email	34	11.41
3	signup_date	6	2.01
4	source	15	5.03
5	region	30	10.07
6	plan_selected	14	4.70
7	marketing_opt_in	0	0.00
8	age	0	0.00
9	gender	14	4.70

- **customer_id**, **age**, and **marketing_opt_in** have no missing values after cleaning.
- Other columns left unfilled **to avoid bias**.
- **email** has the most missing data (11.41%), this is not used in analysis therefore minor.
- **Region** has 10.07% missing data, which should be considered when performing segmentation analysis.

📌 Part 2: Consistent Category Value Checks

This output shows the cleaned distribution of key categorical columns, confirming consistent formatting and highlighting remaining gaps:

```
📊 Totals by Category in 'plan_selected':
plan_selected
Premium    99
Pro        93
Basic      92
NaN        14
Name: count, dtype: int64

📊 Totals by Category in 'gender':
gender
Female        92
Male          91
Other         59
Non-Binary    42
NaN           14
Name: count, dtype: int64

📊 Totals by Category in 'marketing_opt_in':
marketing_opt_in
No     167
Yes    131
Name: count, dtype: int64

📊 Totals by Category in 'region':
region
North      65
East       61
South      58
West       45
Central    39
NaN        30
Name: count, dtype: int64

📊 Totals by Category in 'source':
source
Youtube      58
Google       50
Referral     49
Instagram    48
Facebook     40
Linkedin     38
NaN          15
Name: count, dtype: int64
```

Plan Selected: Labels are standardised; 14 entries still missing.

- **Gender**: Categories are consistently formatted; 14 values remain missing.
- **Marketing Opt-In**: Cleanly split between 'Yes' and 'No'; no missing values.
- **Region**: Contains expected categories; 30 values still missing.
- **Source**: Shows distinct acquisition channels; 15 missing entries remain.

Section 3: User Behavior Summary / Key Findings

Introduction

Now that the data is clean, we explore overall trends in user sign-ups, such as patterns by date, region, source, plan, age, and marketing preferences.

Part 1: Weekly Sign-Ups Summary

- Weekly customer sign-ups are shown from **January to October 2024**.
- Sign-up numbers are consistent, typically **between 6 and 7 per week**.
- Only **minor fluctuations** are observed.
- Indicates a **steady and reliable** customer acquisition rate.

```
📅 Weekly Sign-ups:
signup_week
2024-01-01    6
2024-01-08    7
2024-01-15    7
2024-01-22    7
2024-01-29    8
2024-02-05    7
2024-02-12    7
2024-02-19    7
2024-02-26    7
2024-03-04    7
2024-03-11    7
2024-03-18    6
2024-03-25    6
2024-04-01    7
2024-04-08    7
2024-04-15    7
2024-04-22    7
2024-04-29    6
2024-05-06    7
2024-05-13    7
2024-05-20    7
2024-05-27    7
2024-06-03    6
2024-06-10    6
2024-06-17    7
2024-06-24    7
2024-07-01    7
2024-07-08    7
2024-07-15    6
2024-07-22    7
2024-07-29    7
2024-08-05    6
2024-08-12    7
2024-08-19    7
2024-08-26    7
2024-09-02    7
2024-09-09    7
2024-09-16    7
2024-09-23    7
2024-09-30    7
2024-10-07    7
2024-10-14    6
2024-10-21    6
Name: customer_id, dtype: int64
```

Part 2: Further Summary Statistics

Key Points:

- **Top sources**: YouTube, Google, and Referral.
- **Regions**: Most users are from North and East; 30 unknown entries.
- **Plans**: Premium is the most selected, followed closely by Pro and Basic.
- **Marketing Opt-in**: Opt-out was slightly more common across all genders.
- **Age**: Ranges from 21 to 60, with a mean of 35.47 and no missing values.

These insights help profile customer behaviour and guide future targeting.

```
Sign-ups by Source:
source
Youtube      58
Google       50
Referral     49
Instagram    48
Facebook     40
Linkedin     38
Unknown      15
Name: count, dtype: int64

Sign-ups by Region:
region
North      65
East       61
South      58
West       45
Central    39
Unknown    30
Name: count, dtype: int64

Sign-ups by Plan Selected:
plan_selected
Premium    99
Pro        93
Basic      92
Unknown    14
Name: count, dtype: int64

Marketing Opt-in by Gender:
marketing_opt_in  No  Yes
gender
Female            48   44
Male              54   37
Non-Binary        23   19
Other             35   24
Unknown            7    7

Age Summary Statistics:
    Min   Max   Mean  Median  Nulls
0  21.0  60.0  35.47    34.0      0
```

Section 4: Support Ticket Analysis (Optional)

📌 Introduction

As a stretch task, we load the support ticket dataset and join it to our customer data to assess support behavior. We aim to understand how support activity varies across plans and regions, and how many customers reached out within 2 weeks of signing up.

📌 Support Activity by Plan & Region

To explore support behaviour, we joined the support_tickets.csv dataset with our cleaned customer data using customer_id.

📊 All-Time Support Activity by Plan and Region:

	plan_selected	region	Total_Support_Tickets
0	Basic	Central	2
1	Basic	East	11
2	Basic	North	3
3	Basic	South	14
4	Basic	West	10
5	Premium	Central	6
6	Premium	East	1
7	Premium	North	6
8	Premium	South	2
9	Premium	West	11
10	Pro	Central	10
11	Pro	East	14
12	Pro	North	11
13	Pro	South	3
14	Pro	West	6

Insights:

As plan level increases (from Basic → Premium → Pro), **support activity appears to shift geographically from the South to the North and East**, possibly reflecting regional differences in user needs, expectations, or plan uptake.

Within 2 weeks of Signup:

```
Support Activity within 2 weeks of Signup by Plan and Region:
   plan_selected   region  Support_Tickets
0          Basic     East                1
1          Basic    North                1
2          Basic    South                7
3          Basic     West                4
4        Premium  Central                1
5        Premium    North                1
6        Premium     West                6
7            Pro  Central                2
8            Pro     East                6
9            Pro    North                7
10           Pro    South                2
11           Pro     West                1
              Users_on_Plan  Total_Support_Tickets  Tickets_per_User
```

```
Customers who contacted support within 2 weeks of signup: 29
```

Section 5: Business Insights

📌 Introduction

This section answers key business questions using the cleaned data, providing insights to support marketing, onboarding, and campaign decisions.

📌 Business Questions and Answers

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

Answer:

In September 2024 (the last full-recorded month), the acquisition source that brought in the most users was **YouTube**, with **7 new users** signing up through this channel.

```
Last full month: 2024-09
📊 Acquisition Source Counts for Last Month:
source
Youtube      7
Referral     6
Google       5
Linkedin     5
Facebook     3
Instagram    2
Name: count, dtype: int64

✅ Top acquisition source in 2024-09: 'Youtube' with 7 users
```

---

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

Answer:

Although 'Central' has the lowest count among defined regions, the real sign of incomplete data lies in the **30 unassigned users** with missing region values. This gap reflects a **collection issue** rather than a lack of users from a specific region.

---

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

Answer:

Older users are slightly more likely to opt in to marketing. On average, those who opted in were 0.87 years older than those who did not, though the median age was the same (34) for both groups.

```
📊 Age Summary by Marketing Opt-In:
                  count   mean  median
marketing_opt_in
No                  167  35.08    34.0
Yes                 131  35.95    34.0

📌 On average, users who opted in are older by 0.87 years.
```

---

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

Answer:

The most commonly selected plan is **Premium**, chosen by the highest number of users. Among those who selected the Premium plan, the most common age group is **40 years old**, with **23 users** in that age group. This suggests that users around age 40 are the most engaged with the Premium offering.

```
Most commonly selected plan: Premium
Most common age group for Premium: Age 40

Age distribution for most common plan:
age
21     6
25    16
29    15
34    20
40    23
47     9
53     6
60     4
Name: count, dtype: int64
```

---

5. (Optional) Which plan's users are most likely to contact support?

Answer:

Pro plan users are the most likely to contact support (0.47 tickets per user), suggesting either more advanced feature usage requiring assistance or higher expectations for support responsiveness.

plan_selected	Users_on_Plan	Total_Support_Tickets	Tickets_per_User
Basic	92	40	0.43
Premium	99	26	0.26
Pro	93	44	0.47

Section 6: Business Recommendations, Data Issues and Risks

Introduction

This section provides practical business recommendations based on data trends and highlights key data quality issues. It offers ideas to improve customer engagement and suggests ways to enhance data accuracy in future reporting.

Part 1: Business Recommendations

1. Focus marketing efforts on the 34–40 age group, which showed the highest engagement and uptake of the Premium plan. This group appears most responsive to higher-tier services.

2. Prioritise acquisition via YouTube and Google, as they consistently bring in the most users. Consider increasing ad spend or content on these platforms.

3. Improve overall data collection, enforcing required fields and formatting during signup to ensure consistent and complete entries.

Part 2: Data Issues and Risks

Issue: Only customer_id was used to find duplicates. Duplicate or invalid emails were not checked, so the same person might appear more than once under different IDs.

Risk: This could lead to inflated user counts and misleading insights, especially in campaign tracking or engagement analysis.

Solution: Future checks should include email validation and deduplication using methods like lowercasing, regex, and duplicate detection. This would improve accuracy in tracking and communications.

[END OF PROJECT]