Project Coversheet

Full Name	Polislava Kalcheva
Email	polislavakk@gmail.com
Contact Number	+359884761900
Date of Submission	25/07/2025
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

1. Introduction

Briefly describe the task and dataset:

This project analyses customer sign-up data to understand user behaviour and assess data quality. The goal is to uncover key patterns (e.g., marketing opt-ins, preferred plans, regional trends) and identify issues such as missing or inconsistent data. Two datasets were used: customer_signups.csv and support_tickets.csv.

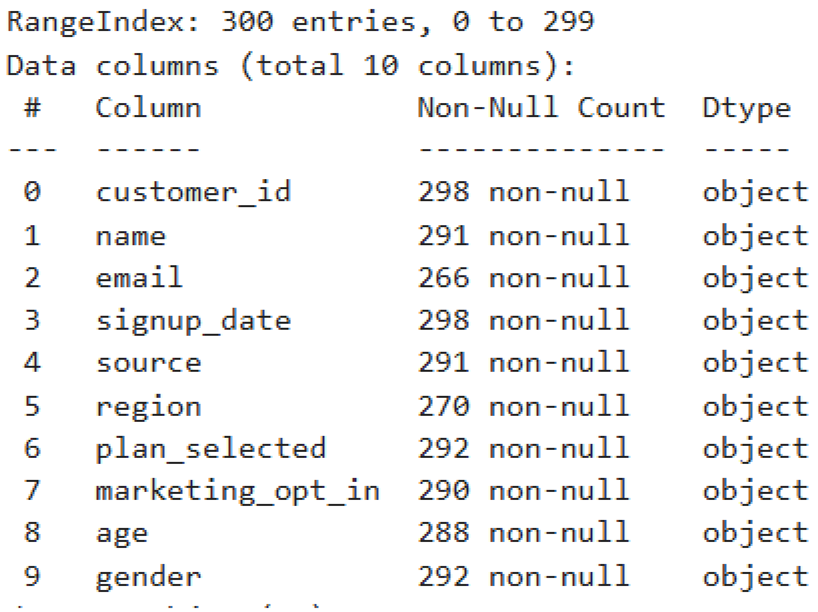
2. Data Cleaning Summary

Explain what you cleaned and how.

Mention duplicates removed, missing data handled, and standardisations made

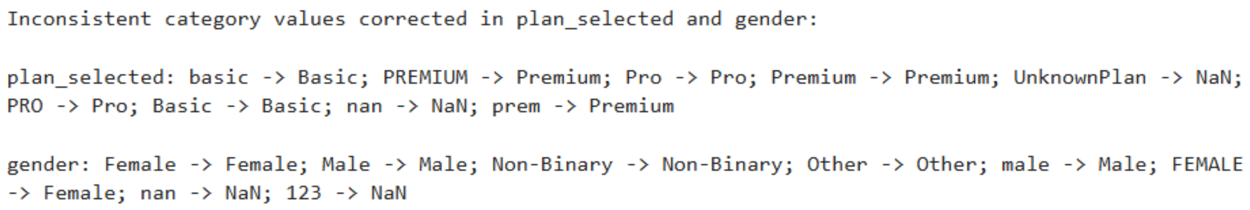
(optional: include screenshot of output)

In the dataset customer_signups.csv there are 300 entries originally and each column has less than 10 missing data inputs.



```
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
```

Moreover, the column ‘signup_date’ was converted into date time type. Then, the column ‘plan_selected’ needed cleaning of the duplicates in the text such as ‘Pro’ and ‘PRO’ to be converted to just ‘Pro’, ‘PREMIUM’ and ‘prem’ to be Premium both, and others, shown below the full transformation per column. In the column ‘gender’ there were irrelevant inputs such as ‘123’ which were converted to empty values (NaN).



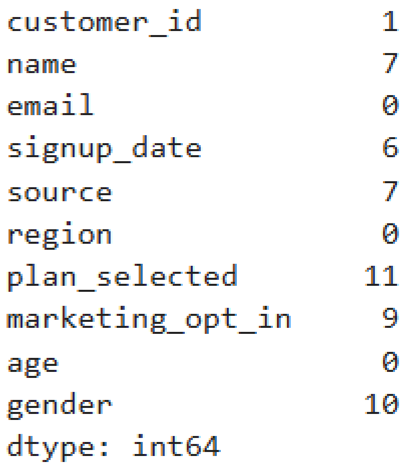
```
Inconsistent category values corrected in plan_selected and gender:

plan_selected: basic -> Basic; PREMIUM -> Premium; Pro -> Pro; Premium -> Premium; UnknownPlan -> NaN;
PRO -> Pro; Basic -> Basic; nan -> NaN; prem -> Premium

gender: Female -> Female; Male -> Male; Non-Binary -> Non-Binary; Other -> Other; male -> Male; FEMALE
-> Female; nan -> NaN; 123 -> NaN
```

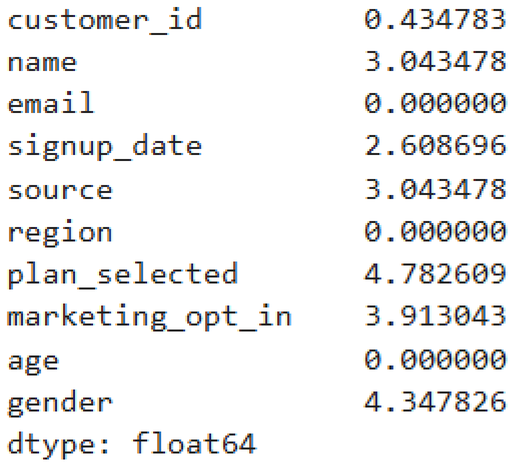
Therefore, in the column ‘customer_id’ there was 1 duplicate which was removed. Then, the missing values were handled in the ‘region’, ‘email’, and ‘age’ column such as removing the rows from the table where any of these columns may have empty value (NaN).

The statistics of the dataset show the remaining empty values per column.



```
customer_id          1
name                 7
email                0
signup_date          6
source               7
region               0
plan_selected       11
marketing_opt_in     9
age                  0
gender              10
dtype: int64
```

Moreover, the percentage of the missing values per columns is shown below, indicating that the ‘plan_selected’ column has the highest empty values followed by the ‘gender’ column.



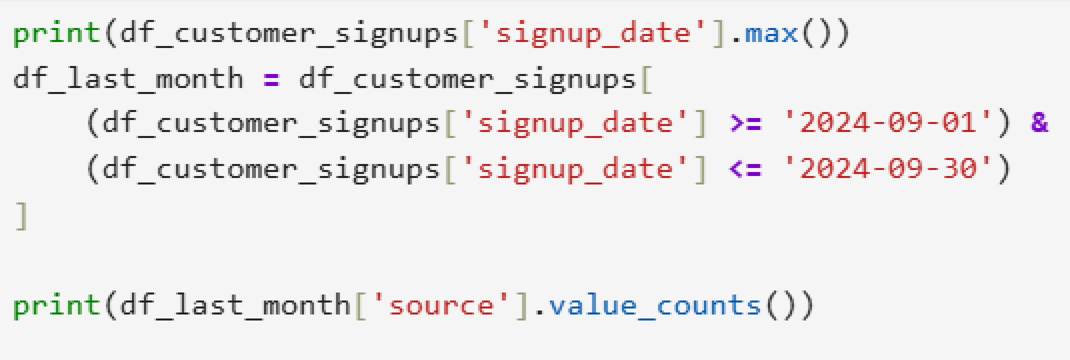
```
customer_id         0.434783
name                3.043478
email               0.000000
signup_date         2.608696
source              3.043478
region              0.000000
plan_selected       4.782609
marketing_opt_in    3.913043
age                 0.000000
gender              4.347826
dtype: float64
```

3. Key Findings & Trends

Write 2-3 short insights based on the outputs

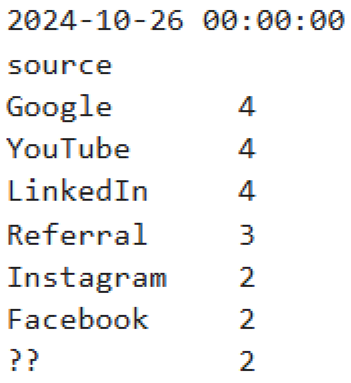
Optional: Include screenshots output

The acquisition sources that brought in the most users in last full month September were Google, YouTube and LinkedIn, with 4 signups each. The rest are Referral with 3, Instagram with 2 and Facebook with 2 and unknown with 2. Check the code and the output below.



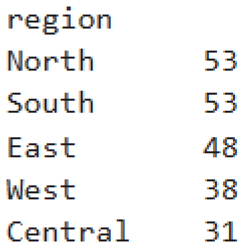
```
print(df_customer_signups['signup_date'].max())
df_last_month = df_customer_signups[
    (df_customer_signups['signup_date'] >= '2024-09-01') &
    (df_customer_signups['signup_date'] <= '2024-09-30')
]

print(df_last_month['source'].value_counts())
```



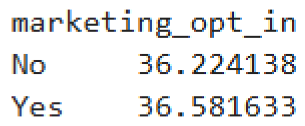
```
2024-10-26 00:00:00
source
Google       4
YouTube      4
LinkedIn     4
Referral     3
Instagram    2
Facebook     2
??           2
```

It was check whether a region shows signs of missing or incomplete data. There are no missing NaN or placeholder ‘??’ values in the region column. However, the ‘Central’ and ‘West’ regions have significantly fewer users compared to others, which may indicate incomplete data from those areas. Check the code output below.



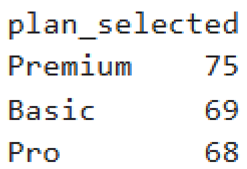
```
region
North      53
South      53
East       48
West       38
Central    31
```

It was checked whether older users more or less likely to opt in to marketing. And the answer is that older users are slightly more likely to opt in to marketing, but the difference in average age of younger and older with 36.58 years and 36.22 years respectively, is minimal and may not be meaningful. Check the code output below.

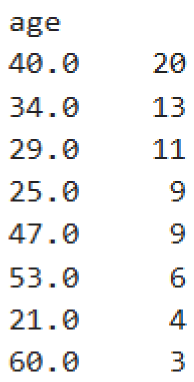


```
marketing_opt_in
No      36.224138
Yes     36.581633
```

There for another insight is that the most commonly selected plan is ‘Premium’ with 75 users in any age from the dataset, and it is the most popular among users aged around 40 years old of 20 users out of the 75 users. Check the code output below.



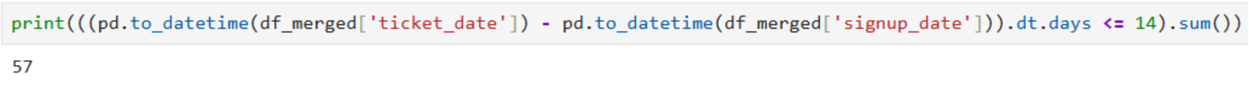
```
plan_selected
Premium    75
Basic      69
Pro        68
```



```
age
40.0    20
34.0    13
29.0    11
25.0     9
47.0     9
53.0     6
21.0     4
60.0     3
```

Another question to answer is ‘Which plan’s users are most likely to contact support?’. This question cannot be answered with the current dataset, as there is no information about user support contact. To assess this, additional data such as support tickets or contact logs would be required. While the dataset does not include direct information on support contact, one might assume that Premium plan users are more likely to contact support. This is because Premium users are likely paying more and may expect more assistance or encounter more features, increasing the chance of needing help.

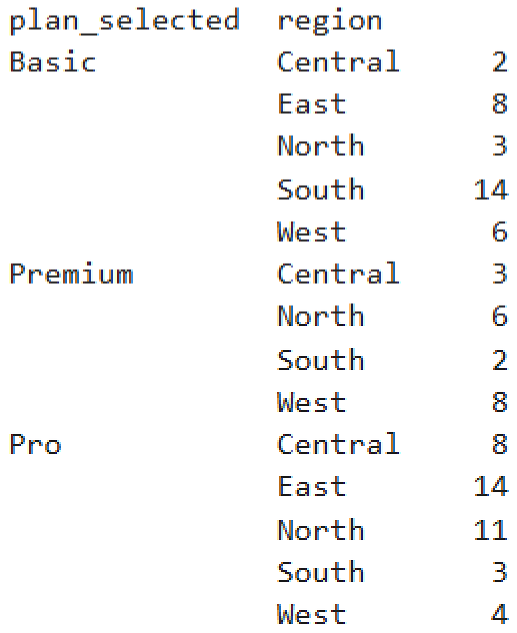
According to the second dataset support_tickets.csv, there are two insights. The first is about how many customers contacted support within 2 weeks of sign-up. They are 57 based on the following code.



```
print(((pd.to_datetime(df_merged['ticket_date']) - pd.to_datetime(df_merged['signup_date'])).dt.days <= 14).sum())

57
```

And the last insight is that the ‘Basic’ plan had the most support requests from the South region followed by East with 14 and 8 respectively. The ‘Pro’ plan also had high support activity from the East and North with 14 and 11 respectively. The ‘Premium’ plan users contacted support the most from the North region with 6, with smaller numbers across other regions.



```
plan_selected  region
Basic          Central     2
               East        8
               North       3
               South      14
               West        6
Premium        Central     3
               North       6
               South       2
               West        8
Pro            Central     8
               East       14
               North      11
               South       3
               West        4
```

4. Business Question Answers

Clearly answer each question with short explanations

5. Recommendations

Suggest 2-3 ideas based on your findings (e.g., focus campaigns on the most engaged age group, improve data collection for missing regions)

Focus on the 30 to 40 age group: Since most Premium plan users are around 40 years old, marketing campaigns should focus on users aged 30 to 40, as they appear the most engaged.

Improve data collection in Central and West regions: These regions had the lowest sign-up numbers. Either fewer users signed up or the data was not fully captured. It would help to ensure proper regional tracking in future forms or systems.

Provide stronger onboarding for ‘Basic’ and ‘Pro’ users: These users had the highest support requests in some regions. Improving the onboarding experience might reduce the number of support tickets.

6. Data Issues or Risks

Highlight one data quality problem

Explain how it could be fixed at the source or in future reporting.

Issue: Some rows had missing or irrelevant symbols as values such as empty value (NaN) or ‘??’, or ‘123’ in important columns such as region, plan_selected, and gender. These had to be removed or cleaned, which reduced the dataset size and may bias the results.

The solution is forms used during sign-up should include validation, for example, making sure the region is selected from a dropdown menu and that only valid options are accepted. Moreover, ensuring that all required fields are filled before submission can reduce missing data.