Project Coversheet

Full Name	Kirana Dhinakar Raj
Email	kiranadhinakar@gmail.com
Contact Number	+49 179 5665425
Date of Submission	31.08.2025
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

CUSTOMER SIGN-UP BEHAVIOR & DATA QUALITY AUDIT

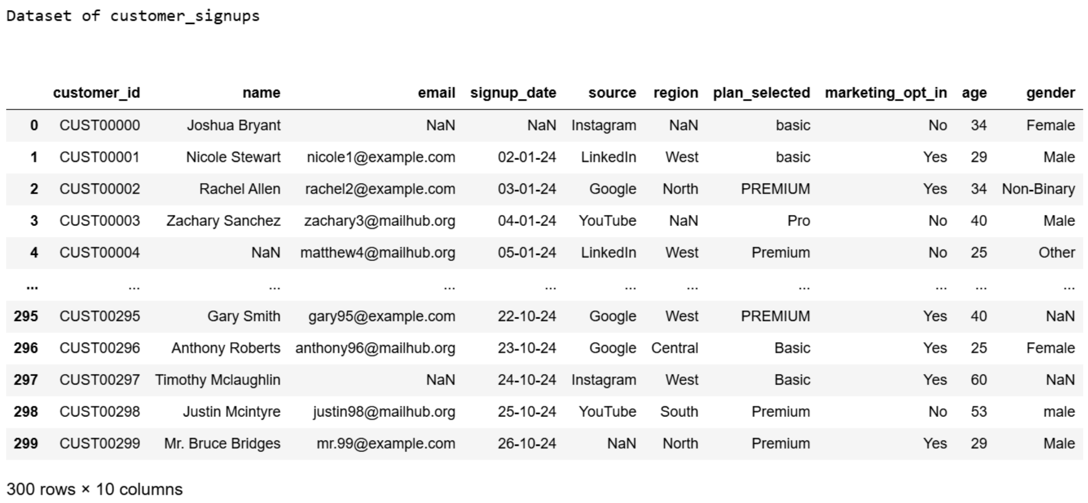
1. INTRODUCTION:

A dataset of customer signup details from a SaaS company offering tiered subscription plans was given. The key task was to analyze user behavior data regarding acquisition trends and to support data quality audit. The report would be presented to the Marketing and Onboarding teams, who are interested in identifying in which regions users data is incomplete, how users are signing up, which type of plans they are choosing and analyzing the behavior of the marketing opt-in characteristics. The dataset was given in a .csv file named ‘customer_signups’ and consisted of 300 entries with 10 demographic columns namely ‘customer_id’, ‘name’, ‘email’, ‘signup_date’, ‘source’, ‘region’, ‘plan_selected’, ‘marketing_opt_in’, ‘age’ and ‘gender’. The given dataset was inconsistent and had a lot of missing fields which needed to be formatted for better analysis. Programming for analysis was done in Python and the IDE used was Jupyter Notebook. The .ipynb file that contains the code has been attached along with this.

2. DATA CLEANING SUMMARY:

The step wise procedure followed for the data cleaning process is given below;

- The first step to the programming part would be to import the necessary libraries. Here I used two python libraries namely pandas and numpy for analysis.
- Next, the .csv file was imported to the notebook and then converted into a dataframe. The dataframe helps to list out the data in a tabular format of rows and columns.

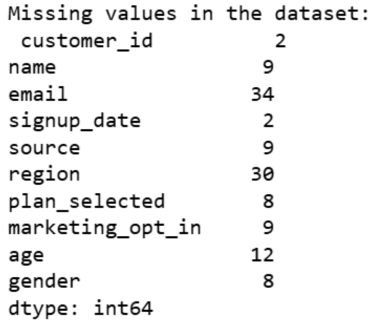


```
Dataset of customer_signups
```

	customer_id	name	email	signup_date	source	region	plan_selected	marketing_opt_in	age	gender
0	CUST00000	Joshua Bryant	NaN	NaN	Instagram	NaN	basic	No	34	Female
1	CUST00001	Nicole Stewart	nicole1@example.com	02-01-24	LinkedIn	West	basic	Yes	29	Male
2	CUST00002	Rachel Allen	rachel2@example.com	03-01-24	Google	North	PREMIUM	Yes	34	Non-Binary
3	CUST00003	Zachary Sanchez	zachary3@mailhub.org	04-01-24	YouTube	NaN	Pro	No	40	Male
4	CUST00004	NaN	matthew4@mailhub.org	05-01-24	LinkedIn	West	Premium	No	25	Other
...	...	...	...	...	...	...	...	...	...	...
295	CUST00295	Gary Smith	gary95@example.com	22-10-24	Google	West	PREMIUM	Yes	40	NaN
296	CUST00296	Anthony Roberts	anthony96@mailhub.org	23-10-24	Google	Central	Basic	Yes	25	Female
297	CUST00297	Timothy Mclaughlin	NaN	24-10-24	Instagram	West	Basic	Yes	60	NaN
298	CUST00298	Justin Mcintyre	justin98@mailhub.org	25-10-24	YouTube	South	Premium	No	53	male
299	CUST00299	Mr. Bruce Bridges	mr.99@example.com	26-10-24	NaN	North	Premium	Yes	29	Male

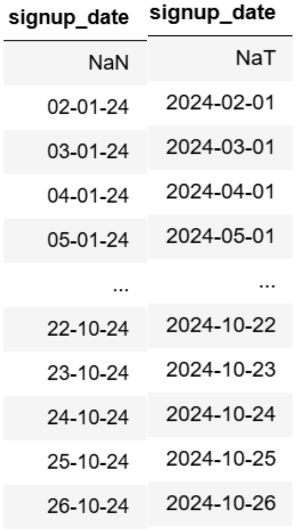
300 rows × 10 columns

- To identify the number of rows missing against each column of the dataset, I used the df.isnull().sum() method. Also, the percentages of those missing values was calculated. All the data was of type string object, which was found out by the df.dtypes command.



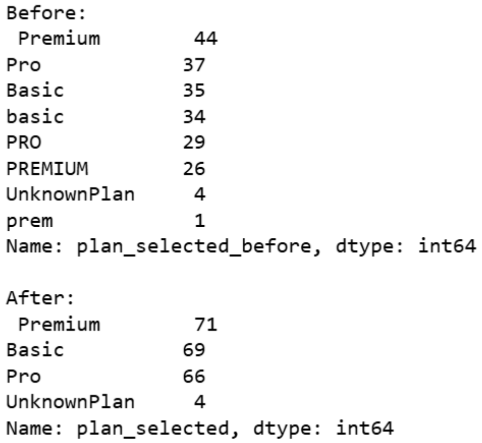
```
Missing values in the dataset:
 customer_id          2
name                 9
email               34
signup_date          2
source               9
region              30
plan_selected        8
marketing_opt_in     9
age                 12
gender               8
dtype: int64
```

- The next step was to identify if there were any duplicate entries in the dataset. Fortunately, there were none.
- The next task was to convert the ‘signup_date’ column into datetime. However, there was an error ‘**ParserError**: Unknown string format: not a date present at position 120’. To avoid this, all indices (rows) where ‘signup_date’ was ‘not a date’ was removed. This enabled me to convert the remaining entries in the ‘signup_date’ to ‘datetime’ format. The screenshots below show the conversion, before on the left side and after on the right side.



signup_date	signup_date
NaN	NaT
02-01-24	2024-02-01
03-01-24	2024-03-01
04-01-24	2024-04-01
05-01-24	2024-05-01
...	...
22-10-24	2024-10-22
23-10-24	2024-10-23
24-10-24	2024-10-24
25-10-24	2024-10-25
26-10-24	2024-10-26

- Presence of missing values in the dataset does not help in analysis. These were handled by filling in the ‘NaN’ entries with blank space and then dropping the rows with empty values in the respective columns, for eg: email, region etc. After this step, the dataframe was consolidated to 210 rows with 10 columns.
- To bring uniformity in the ‘name’ column, I stripped the prefixes which were present in a few using the df.str.strip() command.
- The dataframe had quite a number of inconsistent values in a few columns. For eg; PRO → Pro in ‘plan_selected’ column which were also corrected by replacing with the correct text. The count of the number of users based on the ‘customer_id’ before and after standardizing the values was asked. This was calculated using the .’value_counts’ command and is shown below;



```
Before:
 Premium        44
Pro            37
Basic          35
basic          34
PRO            29
PREMIUM        26
UnknownPlan     4
prem            1
Name: plan_selected_before, dtype: int64

After:
 Premium        71
Basic          69
Pro            66
UnknownPlan     4
Name: plan_selected, dtype: int64
```

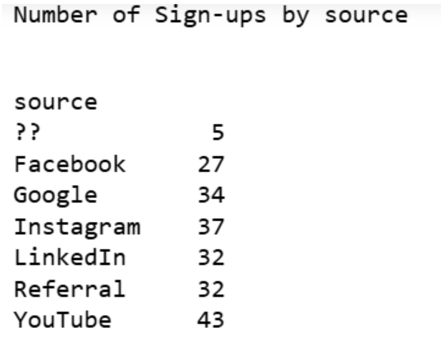
- Similarly, the ‘.value_counts’ command was used for the gender column as well.

- The last step after all cleaning activities, is to ensure that the index values are in order. They were re-ordered using the ‘df.reset_index(drop= True=)’ command.

3. KEY FINDINGS AND TRENDS:

The next task given was to summarize the output findings using pandas aggregation by grouping it based on customer_ids.

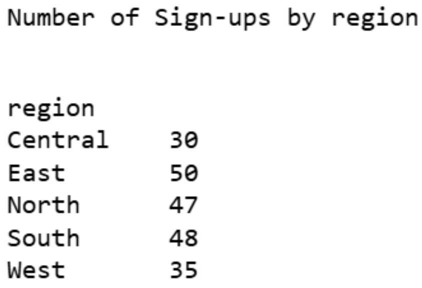
- **Number of customer signups by source:** The findings show that YouTube had the most customer signups in the year 2024 with about 43 users, followed by Instagram with 37 users.



```
Number of Sign-ups by source


source
??            5
Facebook     27
Google       34
Instagram    37
LinkedIn     32
Referral     32
YouTube      43
```

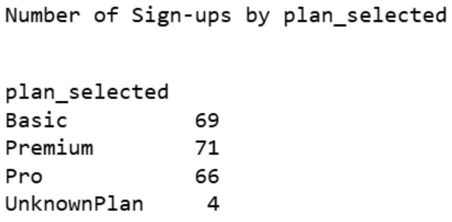
- **Number of customer signups by region:** The data showed the highest number of 50 users signed up from the East region, followed by 48 from the South. The lowest number of users belonged to the Central region.



```
Number of Sign-ups by region


region
Central    30
East       50
North      47
South      48
West       35
```

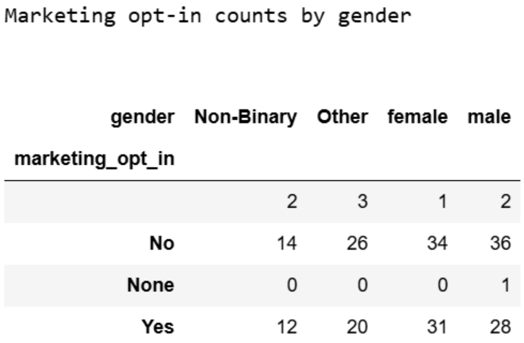
- **Number of signups by plan_selected:** The Premium plans gets the first place with about 71 customers signing up for that. However, here there were 4 users with an unknown plan, which seems like an inconsistency.



```
Number of Sign-ups by plan_selected


plan_selected
Basic          69
Premium        71
Pro            66
UnknownPlan     4
```

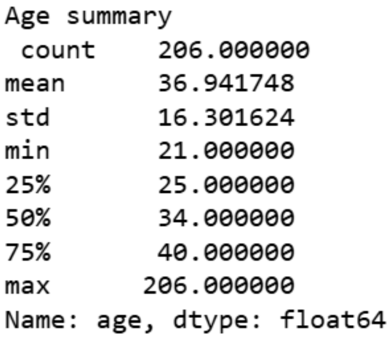
- **Signups per week**: To calculate the number of signups per week, the date was converted to a period of the calendar week and then grouped accordingly. Here the values were almost equally distributed where each week had at least 1 user signing up and a maximum of 8 users.
- **Marketing opt-in counts by gender:** This shows that there were more number of female users than males, opting for a marketing campaign or not.



Marketing opt-in counts by gender

gender	Non-Binary	Other	female	male
marketing_opt_in				
	2	3	1	2
No	14	26	34	36
None	0	0	0	1
Yes	12	20	31	28

- **Age summary**: This was calculated using the '.describe()' method in python as it gives statistics such as count, mean, std, min, max etc.

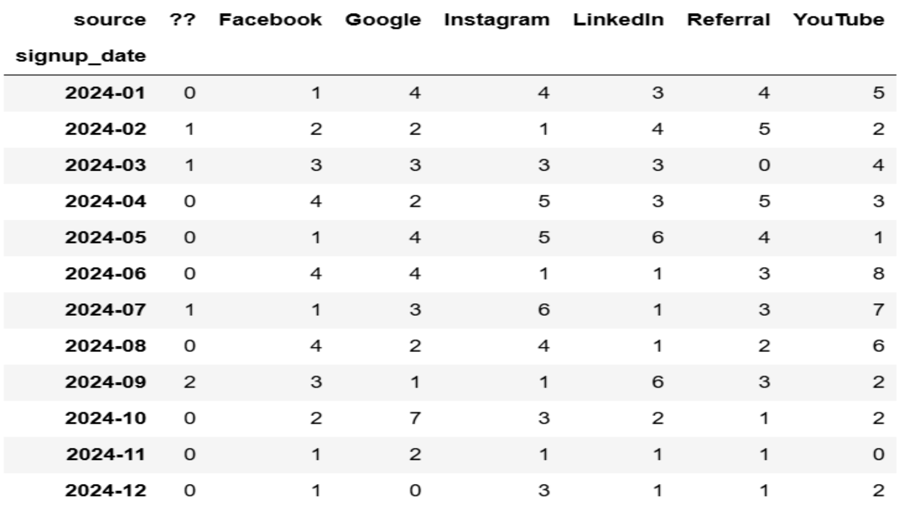


```
Age summary
 count    206.000000
mean      36.941748
std       16.301624
min       21.000000
25%       25.000000
50%       34.000000
75%       40.000000
max      206.000000
Name: age, dtype: float64
```

4. BUSINESS QUESTION ANSWERS:

- Which acquisition source brought in more users last month?

  The analysis shows that in the month of December in 2024, Instagram brought in more number of users with a count of 3, followed by YouTube with 2 users.



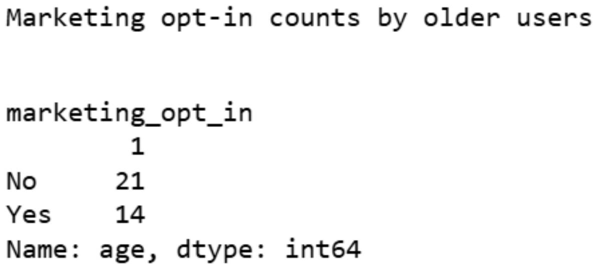
source	??	Facebook	Google	Instagram	LinkedIn	Referral	YouTube
signup_date							
2024-01	0	1	4	4	3	4	5
2024-02	1	2	2	1	4	5	2
2024-03	1	3	3	3	3	0	4
2024-04	0	4	2	5	3	5	3
2024-05	0	1	4	5	6	4	1
2024-06	0	4	4	1	1	3	8
2024-07	1	1	3	6	1	3	7
2024-08	0	4	2	4	1	2	6
2024-09	2	3	1	1	6	3	2
2024-10	0	2	7	3	2	1	2
2024-11	0	1	2	1	1	1	0
2024-12	0	1	0	3	1	1	2

- Which region shows signs of missing or incomplete data?

  All five regions have certain number of data, with the North and West region having the highest number of 11 entries, followed by the South of 9, East of 8 and Central of 4.

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

  To identify this, I split the dataframe into two. One indicating age group of below 50 years and the other indicating age group of more than 50 years of age. The results show that 21 out of 36 users in the age group of above 50, do not want to opt in for marketing and are least interested.

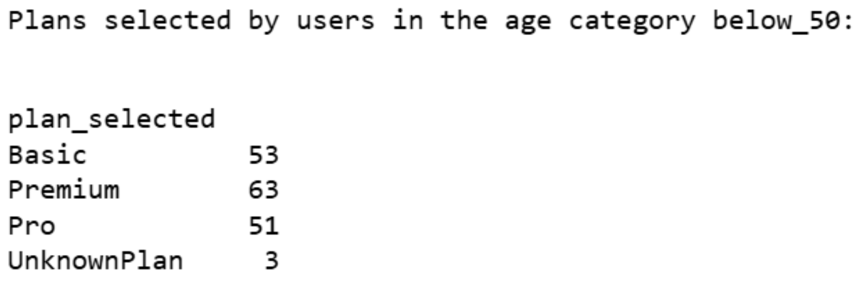


```
Marketing opt-in counts by older users


marketing_opt_in
       1
No     21
Yes    14
Name: age, dtype: int64
```

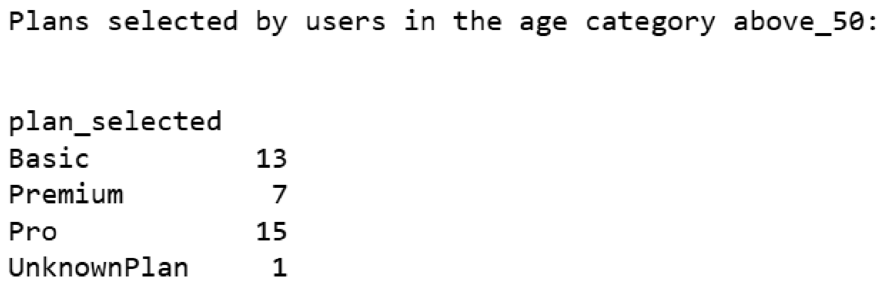
- Which plan is most commonly selected, and by which age group?

  The Premium plan is mostly chosen by the younger age group whereas the older age group of people prefer the Pro plan.



```
Plans selected by users in the age category below_50:


plan_selected
Basic          53
Premium        63
Pro            51
UnknownPlan     3
```



```
Plans selected by users in the age category above_50:


plan_selected
Basic          13
Premium         7
Pro            15
UnknownPlan     1
```

5. RECOMMENDATIONS:

- **Focus campaigns on the younger age group with Premium plans:** Marketing campaigns should highlight premium benefits for younger customers to attract more users in future and also do the same with the Pro plan to the older customer group.
- **Improve data collection for missing regions:** Quite a number of missing entries were seen in the dataset, which should be the focus for improving the analysis and results.

6. DATA ISSUES OR RISKS:

The dataset had a lot of missing, inconsistent data as well as data that was not standardized at the point of collection. This is something that must be taken into account in future to ensure consistency and better analysis performance.