Project Coversheet

Full Name	Shabnam Shaik
Email	shabnamsmiles@gmail.com
Contact Number	+44 7880803036
Date of Submission	28-07-2025
Project Week	Week 1

1. Introduction:

- ### Project Overview

Rapid Scale is a fast-growing SaaS (Software as a Service) company that offers tiered subscription plans tailored to a diverse customer base. The platform focuses on delivering scalable, cloud-based solutions to businesses seeking flexibility, performance, and ease of use.

This project involves analysing the latest customer sign-up dataset to support the **Monthly Business Review (MBR)** and focuses on two primary objectives:

1. **Data Quality Audit** – Assessing the integrity, completeness, and consistency of customer data.
2. **User Acquisition Insights** – Identifying trends in customer sign-ups, preferences, and demographics.

The analysis aims to ensure the integrity and consistency of the data while uncovering patterns in customer sign-ups, preferences, and demographics. The findings will directly support the **Marketing** and **Onboarding** teams in optimizing their campaigns and engagement strategies, ultimately enhancing customer experience and business growth.

- **Data Sources:**

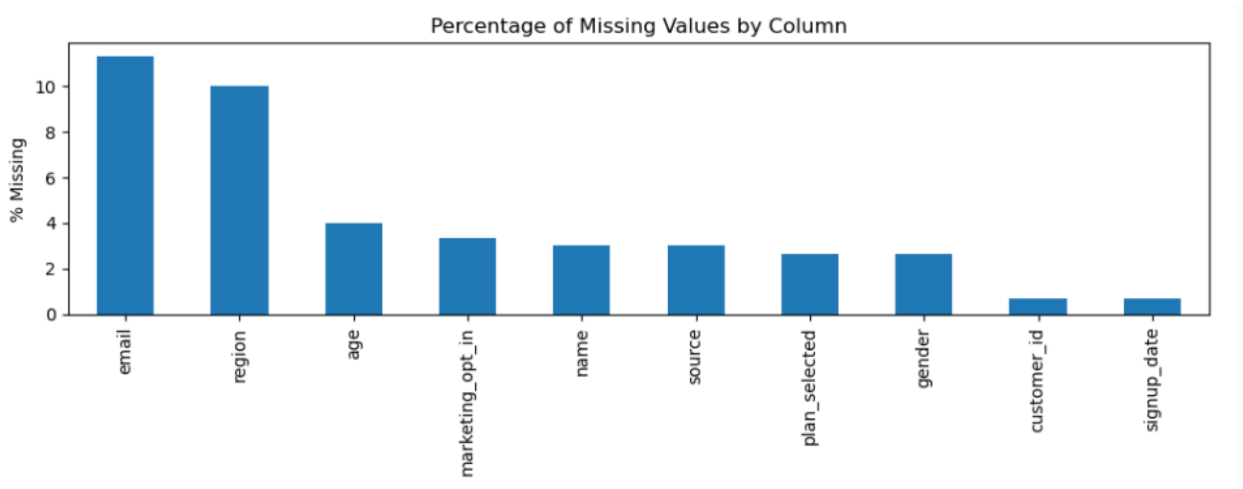
1. customer_signup.csv – Contains customer profile and sign-up related information such as personal details, signup source, and marketing preferences.
2. support_tickets.csv – Contains records of customer support interactions, including ticket metadata and resolution status.

1.a customer_signups.csv – Column Details

Field	Description
customer_id	Unique identifier assigned to each customer in the system.
Name	Full name of the customer provided at signup.
Email	Customer's email address used for communication
signup_date	Date when the customer registered or signed up.
Source	Marketing or acquisition channel that led to the signup (e.g., Google, Instagram, Referral)
Region	Geographic region associated with the customer (Central, North, East, West, South)
plan_selected	The subscription or service plan chosen by the customer during signup. (Basic, Pro, Premium)
marketing_opt_in	Indicates whether the customer agreed to receive marketing emails or promotions (Yes / No / Unknown).
Age	Age of the customer at the time of signup.
gender	Gender of the customer as provided or selected during signup.

1.b customer_signups.csv -- Data Quality Overview

Column	Non-Missing entries	Missing entries	% Missing
customer_id	298	2	0.67%
Name	291	9	3.00%
email	266	34	11.33%
signup_date	298	2	0.67%
Source	291	9	3.00%
region	270	30	10.00%
plan_selected	292	8	2.67%
marketing_opt_in	290	10	3.33%
age	288	12	4.00%
gender	292	8	2.67%



2.a support_tickets.csv – Column Details

Field	Description
ticket_id	Unique identifier for each support ticket raised by a customer.
customer_id	Identifier linking the ticket to the corresponding customer in the system.
ticket_date	Date when the support ticket was created or submitted.
issue_type	Category or type of issue reported (e.g., Billing, Technical, Account Setup,Login Issue, Other).
resolved	Indicates whether the issue has been resolved (Yes / No).

2.b support_tickets.csv -- Data Quality Overview

Column	Non-Missing Entries	Missing Entries	% Missing
ticket_id	123	0	0.00%
customer_id	123	0	0.00%
ticket_date	123	0	0.00%
issue_type	123	0	0.00%
Resolved	123	0	0.00%

2. Data Cleaning Summary

2.1 customer_signups.csv

Field	Data Cleaning steps taken
customer_id	1. Dropped blank rows as it is the primary key 2. No Duplicates were found in this field
Name	Blank values are there but no changes made to this
email	Replaced the blank values of email with 'unknown@example.com'.
signup_date	1. Converted the object datatype to datetime 2. Text format of the date was converted to number format for eg., 2nd February 2024 to 2024-02-02 3. Replaced the 2 blank values of dates with Median of date
Source	Replaced '??' and blank values with 'Unknown'
region	Replaced the blank values with "Unknown" to retain rows
plan_selected	1.Made the data consistent by changing Prem to Premium 2.Replaced blank values with 'Unknownplan'
marketing_opt_in	1.Replaced 'Nil' values with 'No' 2. Replaced blank values with 'Unknown'
age	1. Replaced 'unknown' values with nulls to support age group analysis. 2.Also replaced string format of number to numeric for eg: 'thirty' to 30. 3.Outliers (age >100) are replaced with null values.
gender	1.Standardized values by Replacing 'Non-Binary' to 'Other'. 2. Replaced missing values with "Unspecified"

2.1a Data Status Post-Cleaning

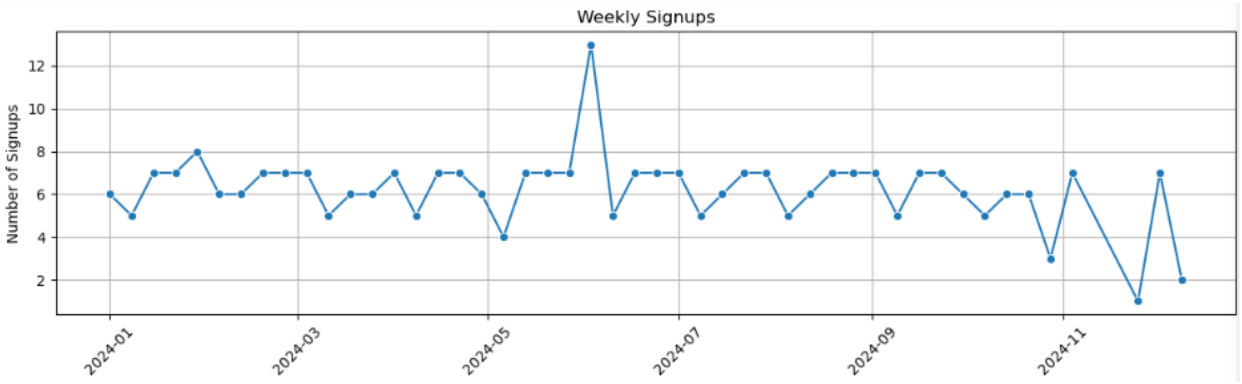
After cleaning and handling null values, the customer signup dataset now has the following data completeness:

Column	Non-Null Count	Data Type	Notes
customer_id	298	object	Complete
Name	289	object	Some missing entries remain
Email	298	object	Complete
signup_date	298	datetime	Complete and converted to date format
Source	298	object	Complete
Region	298	object	Missing values filled or corrected
plan_selected	298	object	Complete
marketing_opt_in	298	object	Complete
Age	279	Int64	Some missing values remain
Gender	298	object	Complete

- **Summary:**

  - Most columns have no missing values after cleaning.
  - Columns name and age still have some missing entries that could not be filled.
  - signup_date has been converted to proper datetime format for accurate analysis.
  - **No data cleaning was required for support_tickets.csv** as the file was already complete and free from missing or inconsistent values upon initial inspection.

3. Key Findings & Trends - -

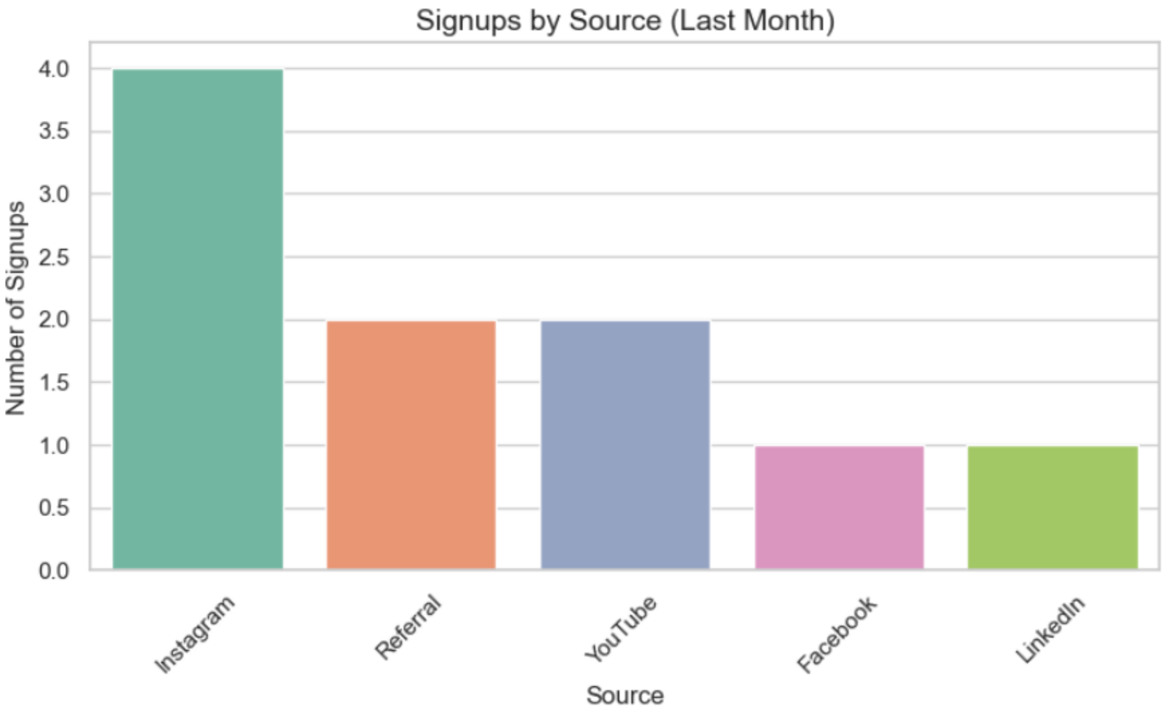


Weekly Signups Summary

Metric	Value
Total Signups	298
Total Weeks Covered	48
Average Weekly Signups	6.2
Highest Signups in a Week	13 (Week of 2024-06-03)
Lowest Signups in a Week	1 (Week of 2024-11-25)
Most Common Weekly Value	7 signups (in 18 separate weeks)
Weeks Below Average (<6)	14
Weeks Above Average (>6)	20
Stable Weeks (6–7 signups)	Majority of the year

4. Business Questions

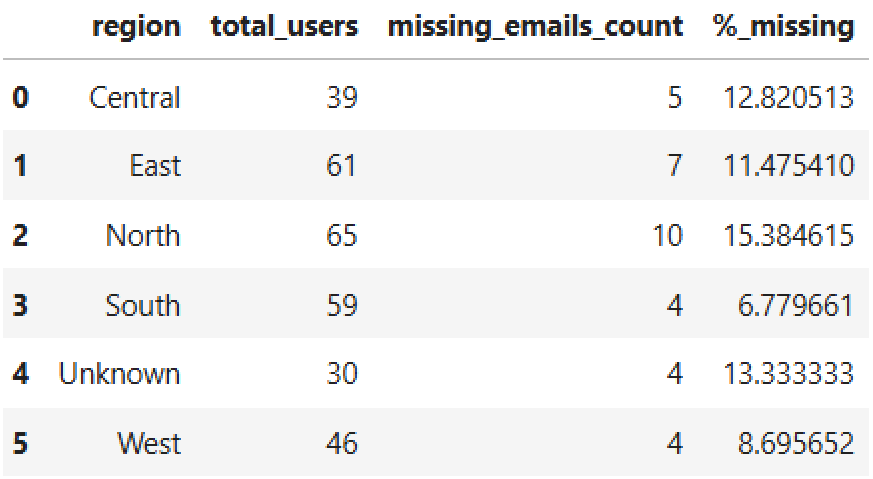
4.1. Which acquisition source brought in the most users last month?



Top acquisition source in December 2024:
Instagram with 4 sign-ups

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

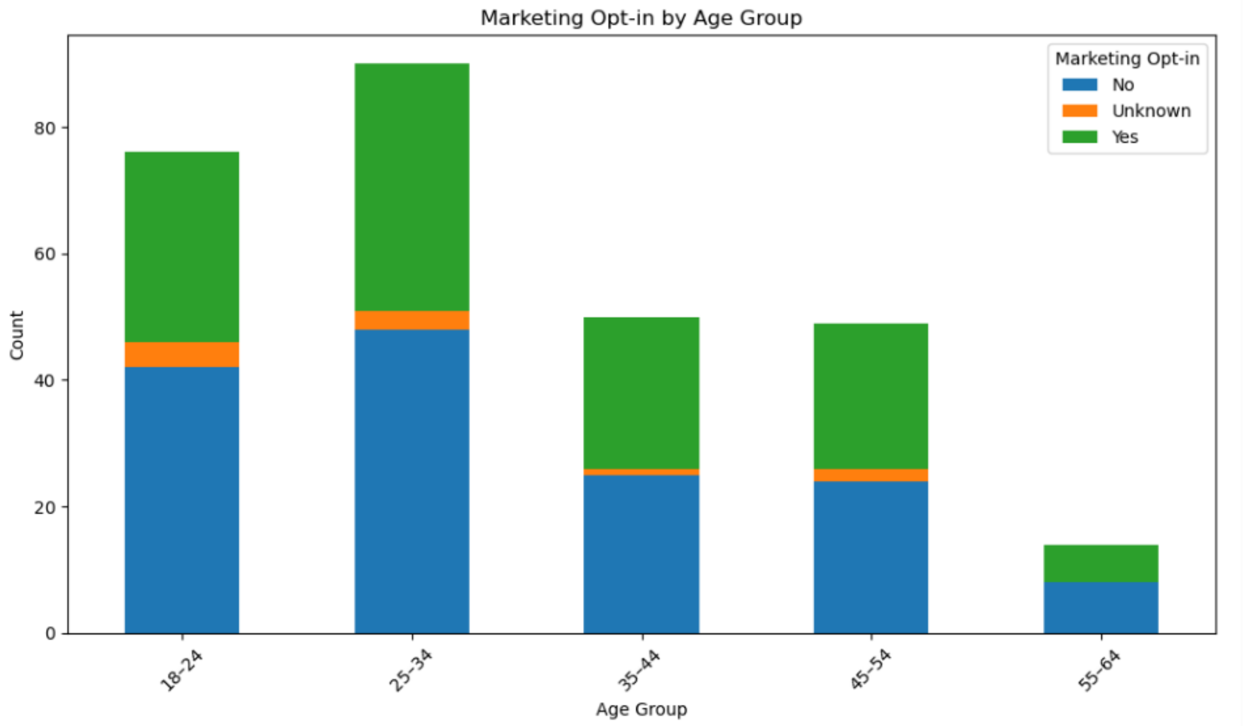
North region shows more missing email Ids, which is the most important field for analysis. Given below are the % details of the missing emails.
The null values in the region are replaced with 'Unknown' to include their counts in the analysis



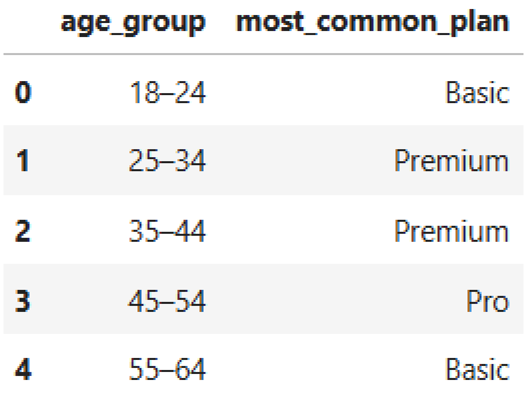
	region	total_users	missing_emails_count	%_missing
0	Central	39	5	12.820513
1	East	61	7	11.475410
2	North	65	10	15.384615
3	South	59	4	6.779661
4	Unknown	30	4	13.333333
5	West	46	4	8.695652

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

` From the below chart, in all the age-groups, users are less likely to opt in to marketing



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

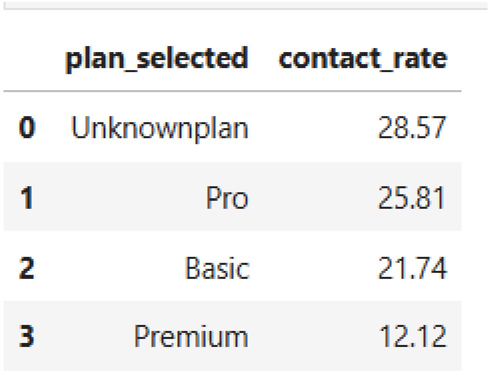


	age_group	most_common_plan
0	18–24	Basic
1	25–34	Premium
2	35–44	Premium
3	45–54	Pro
4	55–64	Basic

4.5. (Optional) Which plan’s users are most likely to contact support?

29% of users who are not under any plan, are the most ones to contact support followed by ‘Pro’ users with 26%

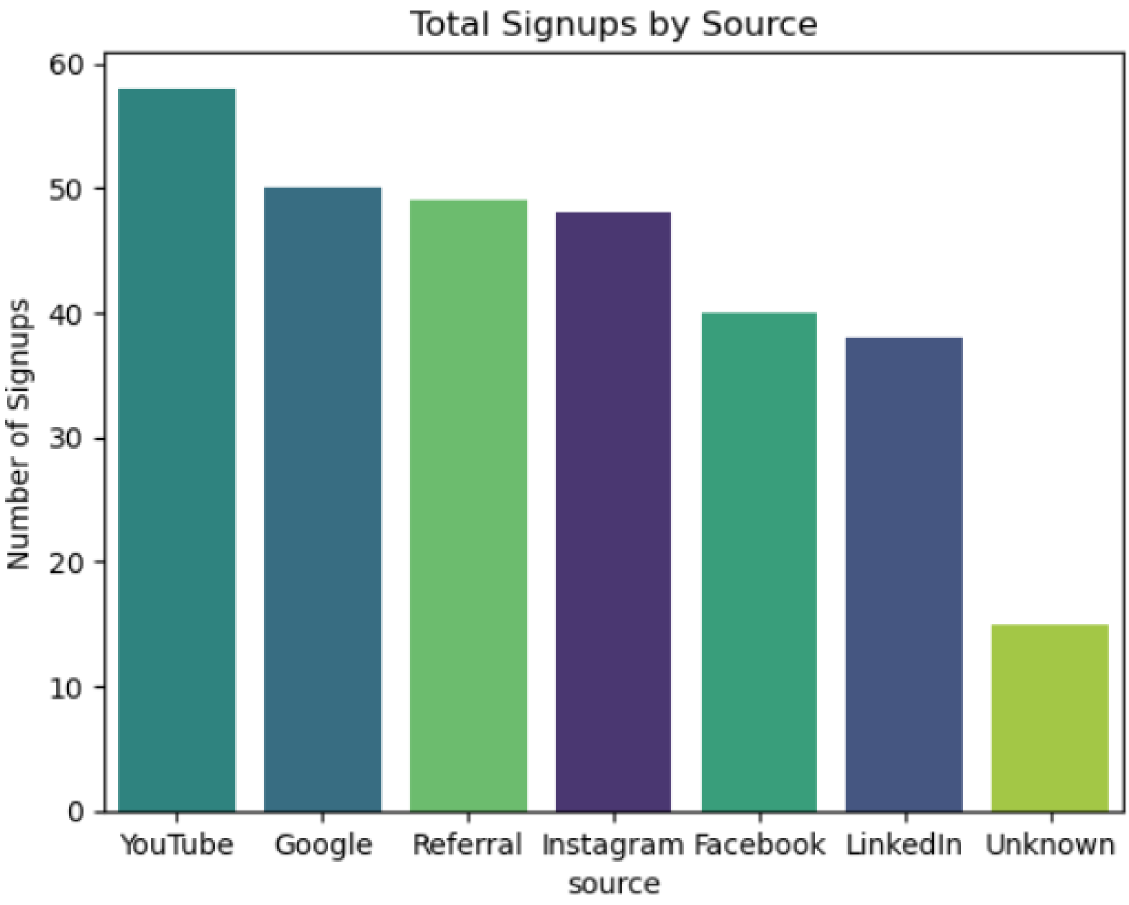
Below details for reference



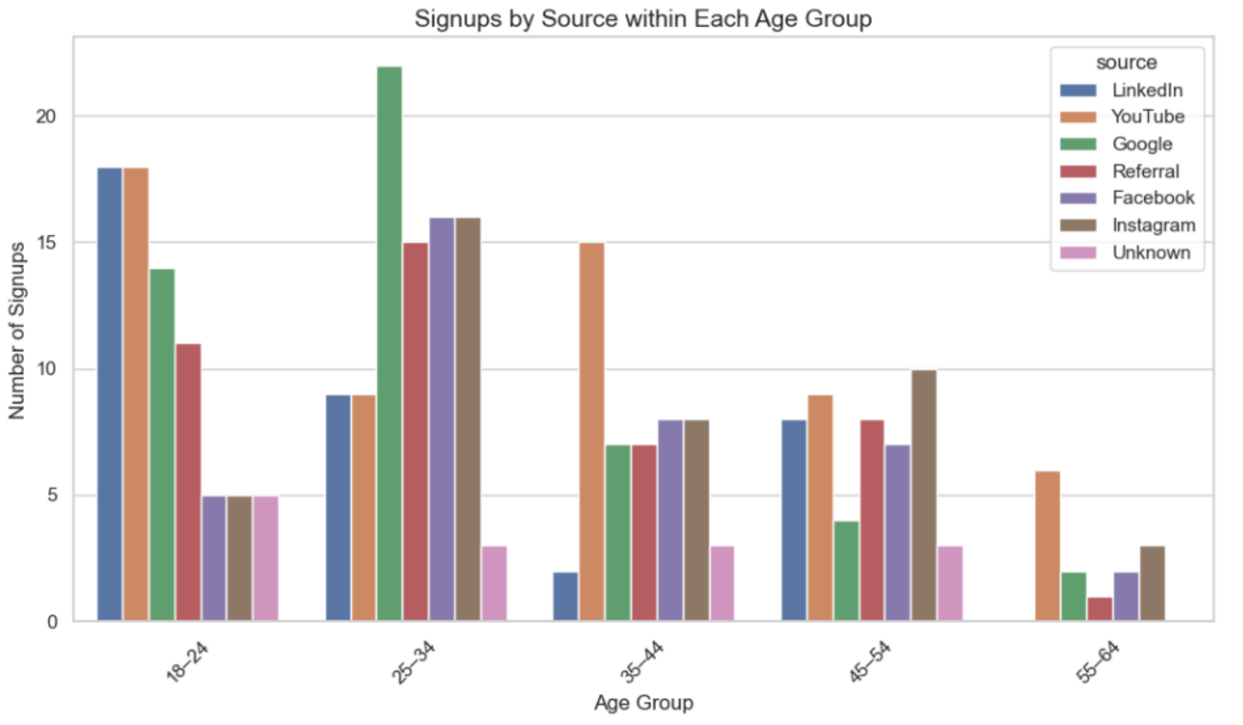
	plan_selected	contact_rate
0	Unknownplan	28.57
1	Pro	25.81
2	Basic	21.74
3	Premium	12.12

5. Recommendations –

5.1 Chart 1: Total Signups by source



5.2. Signups by Source within Each Age Group

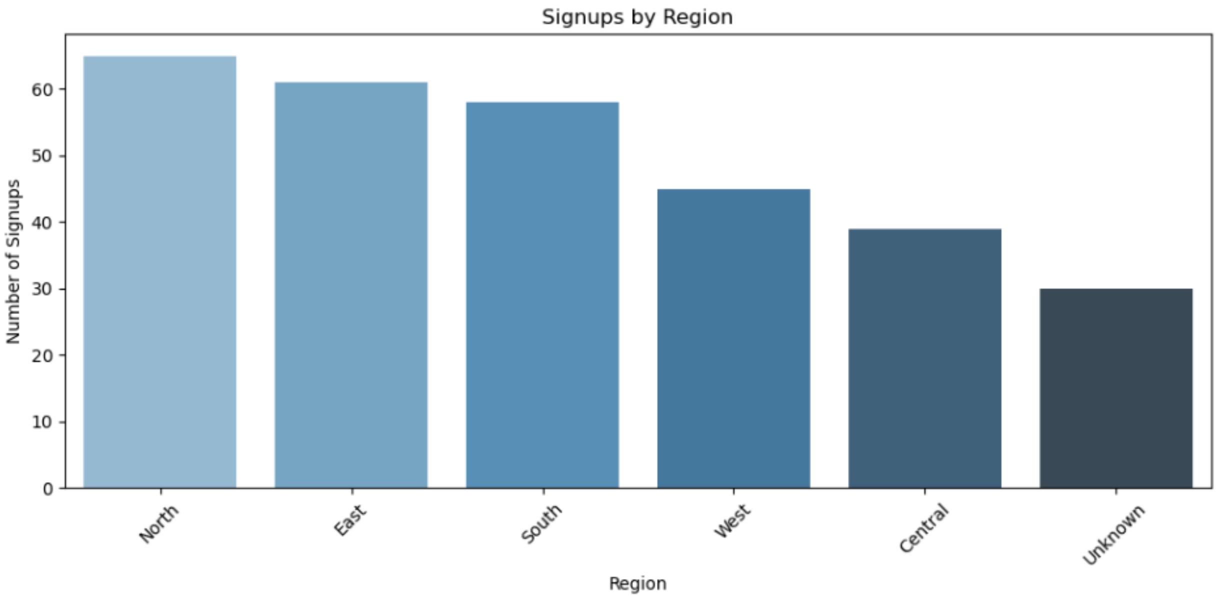


Summary of Signup Sources by Age Group:

While **YouTube is the leading source of signups overall**, the preferred signup source **varies across age groups**. For example:

- **25–34** age group: Highest signups from **Google**
- **18–24** age group: Signups are more evenly distributed, with **LinkedIn** and **YouTube**
- Other age groups also show differing platform preferences
- These insights highlight the importance of **age-targeted marketing**, allowing campaigns to be optimized by focusing on the most effective platforms for each demographic.

5.3. Signups by Region



- Missing values in the region leads to incorrect analysis by region as it is accounting to 10% of the Total values of Region.
- If the region is collected via a form (web, app, etc.), making it **mandatory** to fill before submission might help.
- Use validation to prevent blank or invalid entries.

6. Data Issues or Risks

Despite successful data cleaning and transformation, the following issues and risks were observed:

Issue	Impact
Missing Email Addresses	Prevents direct communication with users; reduces effectiveness of email marketing campaigns.
Missing Names	Limits ability to personalize messages, which may lower engagement (e.g., greeting users by name).
Missing Age Data	Hinders accurate demographic segmentation and age-based targeting in marketing strategies.
Missing Marketing Opt-in	May result in users being excluded from outreach or contacted without clear consent.
Incomplete Regional Data	Affects geographic segmentation and regional campaign analysis.

- ### Project Conclusion

  - This project involved analysing the customer signup dataset for data quality and behavioural insights. The primary objectives were:

    - **Conducting a data quality audit** to assess and improve completeness and consistency
    - **Identifying user acquisition trends** by source, age group, region, and Marketing Opt-in.
    - **Visualizing sign-up patterns** over time to assist in campaign planning and decision-making

  - Post-cleaning, the dataset is now well-structured and reliable for downstream analysis.
  - Key insights such as source popularity by age group and weekly sign-up trends can directly support marketing, product, and onboarding strategies.
  - Future recommendations include enhancing source data validation at the point of entry and reducing optional blanks for key fields (e.g., email, age).